

EUROPEAN PATENT OFFICE

Patent Abstracts of Japan

PUBLICATION NUMBER : 05046324
PUBLICATION DATE : 26-02-93

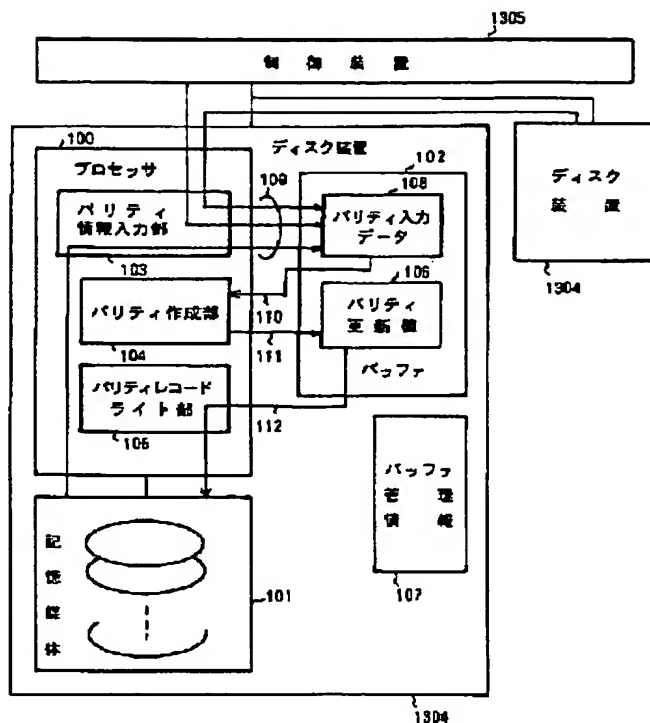
APPLICATION DATE : 20-08-91
APPLICATION NUMBER : 03207808

APPLICANT : HITACHI LTD;

INVENTOR : INOMATA HIROFUMI;

INT.CL. : G06F 3/06 G06F 11/10 G06F 12/16
G06F 12/16

TITLE : STORAGE DEVICE AND STORAGE
DEVICE SYSTEM



ABSTRACT : PURPOSE: To curtail the utilization factor of a data transfer line of a controller and a disk device by dispersing a generating function of an updating value of a parity record to the disk device side, since updating of the parity record becomes overhead, comparing with a conventional processing, in the case a disk array is used.

CONSTITUTION: From a parity information input part 103 in a processor 100, a controller 1305, or a recording medium, information for generating an updating value of a parity record is inputted to a buffer. A parity generating part 104 generates the updating value of the parity record as a parity updating value 106 from the information inputted from the buffer, and stores it in the buffer. Thereafter, by a parity record write part 105, the parity updating value 106 stored in the buffer is written in a recording medium.

COPYRIGHT: (C)1993,JPO&Japio

BEST AVAILABLE COPY

(19)日本国特許庁(JP)

(12) 公開特許公報(A)

(11)特許出願公開番号

特開平5-46324

(43)公開日 平成5年(1993)2月26日

(51)Int.Cl. ⁵	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 3/06	3 0 4 E	7165-5B		
11/10	3 2 0 F	7832-5B		
12/16	3 1 0 J	7629-5B		
	3 2 0 H	7629-5B		

審査請求 未請求 請求項の数7(全 32 頁)

(21)出願番号 特願平3-207808

(22)出願日 平成3年(1991)8月20日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 山本 彰

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(72)発明者 田宮 敏彦

神奈川県小田原市国府津2880番地 株式会社日立製作所小田原工場内

(72)発明者 ▲高▼松 久司

神奈川県小田原市国府津2880番地 株式会社日立製作所小田原工場内

(74)代理人 弁理士 小川 勝男

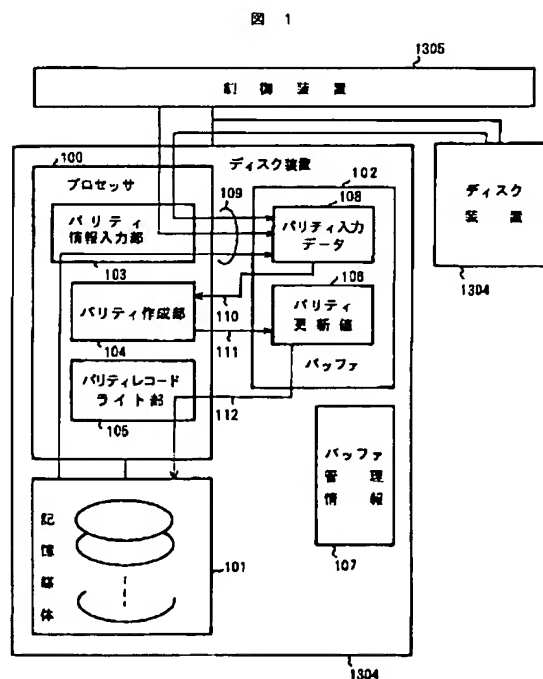
最終頁に続く

(54)【発明の名称】 記憶装置および記憶装置システム

(57)【要約】

【目的】ディスクアレイを用いた場合、パリティレコードの更新が従来处理に比較し、オーバヘッドとなる。このパリティレコードの更新値の作成機能を、ディスク装置側に分散させることにより、制御装置とディスク装置のデータ転送路の利用率を削減する。

【構成】プロセッサ内のパリティ情報入力部、制御装置、あるいは、記録媒体から、パリティレコードの更新値を作成するための情報を、バッファに入力する。パリティ作成部は、バッファに入力した情報より、パリティレコードの更新値をパリティ更新値として作成し、バッファに格納する。この後、パリティレコードライト部により、バッファに格納したパリティ更新値を記録媒体に書き込む。



【特許請求の範囲】

【請求項1】それぞれのレコードグループが、通常のデータを格納した m 個 ($m \geq 1$) のデータレコードと前記通常データを回復するための冗長データを格納した n 個 ($n \geq 1$) の冗長レコードから構成され、少なくとも1つ以上の前記レコードグループに属する1つ以上の前記データレコード、あるいは、1つ以上の前記冗長レコードを格納した記憶媒体を含む記憶装置であって、前記冗長レコードに格納する前記冗長データを作成するための情報を、制御装置、あるいは、前記記憶媒体から受け取る手段と、前記冗長レコードに格納する前記冗長データを作成する手段を有することを特徴とする記憶装置。

【請求項2】それぞれのレコードグループが、通常のデータを格納した m 個 ($m \geq 1$) のデータレコードと前記通常データを回復するための冗長データを格納した n 個 ($n \geq 1$) の冗長レコードから構成され、少なくとも1つ以上の前記レコードグループに属する1つ以上の前記データレコード、あるいは、1つ以上の前記冗長レコードを格納した記憶媒体を含む1つ以上の記憶装置と制御装置を含む記憶システムであって、前記制御装置が、前記データレコードの更新前の値と更新後の値から、前記冗長レコードの更新値を作成するための中間情報を作成する手段と、前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置に作成した前記中間情報を送信する手段を有し、前記記憶装置が、前記制御装置から前記中間情報を受け取る手段と、前記記憶媒体から前記冗長レコードの更新前の値を読み出す手段と、前記中間情報と前記冗長レコードの更新前の値から、前記冗長レコードの更新値を作成する手段を有することを特徴とする記憶装置システム。

【請求項3】それぞれのレコードグループが、通常のデータを格納した m 個 ($m \geq 1$) のデータレコードと前記通常データを回復するための冗長データを格納した n 個 ($n \geq 1$) の冗長レコードから構成され、少なくとも1つ以上の前記レコードグループに属する1つ以上の前記データレコード、あるいは、1つ以上の前記冗長レコードを格納した記憶媒体を含む1つ以上の記憶装置と制御装置を含む記憶システムであって、前記制御装置が、前記パリティグループ内の前記データレコードの更新値の集合を、前記パリティグループ内の前記データレコード、あるいは、前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置にブロードキャストする手段を有し、前記データレコードを格納した前記記憶媒体を含む前記記憶装置が、前記制御装置から前記パリティグループ内の前記データレコードの更新値の集合を受け取る手段と、前記パリティグループ内の前記データレコードの更新値の集合から、前記記憶媒体に書き込むべき前記データレコードの更新値を選択し、前記記憶媒体に書き込む手段と、前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置が、前記制御装置から前記パリティグ

ループ内の前記データレコードの更新値の集合を受け取る手段と、前記パリティグループ内の前記データレコードの更新値の集合から、前記冗長レコードの更新値を作成する手段を有することを特徴とする記憶装置システム。

【請求項4】それぞれのレコードグループが、通常のデータを格納した m 個 ($m \geq 1$) のデータレコードと前記通常データを回復するための冗長データを格納した n 個 ($n \geq 1$) の冗長レコードから構成され、少なくとも1つ以上の前記レコードグループに属する1つ以上の前記データレコード、あるいは、1つ以上の前記冗長レコードを格納した記憶媒体を含む1つ以上の記憶装置を、接続した制御装置であって、前記制御装置が、前記データレコードの更新値を、前記データレコードを格納した前記記憶媒体を含む前記記憶装置と、前記データレコードが属する前記パリティグループ内の前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置に、ブロードキャストする手段を有し、前記データレコードを格納した前記記憶媒体を含む前記記憶装置が、前記制御装置から前記データレコードの更新値を受け取る手段と、前記データレコードの更新前の値を、前記データレコードが属する前記パリティグループ内の前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置に、送信する手段と、前記データレコードの更新値を前記記憶媒体に書き込む手段とを有し、前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置が、前記制御装置から前記データレコードの更新値を受け取る手段と、前記データレコードを格納した前記記憶媒体を含む前記記憶装置から、前記データレコードの更新前の値を、受け取る手段と、前記記憶媒体から前記冗長レコードの更新前の値を読み出す手段と、前記データレコードの更新値と前記データレコードの更新前の値と、前記冗長レコードの更新前の値から、前記冗長レコードの更新値を作成する手段を有することを特徴とする記憶装置システム。

【請求項5】それぞれのレコードグループが、通常のデータを格納した m 個 ($m \geq 1$) のデータレコードと前記通常データを回復するための冗長データを格納した n 個 ($n \geq 1$) の冗長レコードから構成され、少なくとも1つ以上の前記レコードグループに属する1つ以上の前記データレコード、あるいは、1つ以上の前記冗長レコードを格納した記憶媒体とバッファを含む1つ以上の記憶装置を、接続した制御装置であって、前記制御装置が、前記データレコードの更新値を、前記データレコードを格納した前記記憶媒体を含む前記記憶装置と、前記データレコードが属する前記パリティグループ内の前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置に、ブロードキャストする手段を有し、前記データレコードを格納した前記記憶媒体を含む前記記憶装置が、前記制御装置から前記データレコードの更新値を受け取る手段と、前記データレコードが属する前記パリティグ

ープ内の前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置の要求にしたがって、前記データレコードの更新前の値を、前記データレコードが属する前記パリティグループ内の前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置に、送信する手段とを有し、前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置が、前記制御装置から前記データレコードの更新値を受け取り、バッファに格納する手段と、前記バッファに存在しなかった前記データレコードを格納した前記記憶媒体を含む前記記憶装置から、前記データレコードの更新前の値を受け取り、前記バッファに格納する手段と、前記記憶媒体から前記冗長レコードの更新前の値を読みだし、前記バッファに格納する手段と、前記バッファに格納した、前記データレコードの更新値と前記データレコードの更新前の値と、前記冗長レコードの更新前の値から、前記冗長レコードの更新値を作成する手段を有することを特徴とする記憶装置システム。

【請求項6】それぞれのレコードグループが、通常のデータを格納した m 個 ($m \geq 1$) のデータレコードと前記通常データを回復するための冗長データを格納した n 個 ($n \geq 1$) の冗長レコードから構成され、少なくとも1つ以上の前記レコードグループに属する1つ以上の前記データレコード、あるいは、1つ以上の前記冗長レコードを格納した記憶媒体を含む記憶装置であって、制御装置からのリード要求にしたがって、前記データレコードの値を、少なくとも、前記制御装置と、前記データレコードが属する前記パリティグループ内の前記冗長レコードを格納した前記記憶媒体を含む前記記憶装置に、ブロードキャストする手段を有することを特徴とした記憶装置。

【請求項7】それぞれのレコードグループが、通常のデータを格納した m 個 ($m \geq 1$) のデータレコードと前記通常データを回復するための冗長データを格納した n 個 ($n \geq 1$) の冗長レコードから構成され、少なくとも1つ以上の前記レコードグループに属する1つ以上の前記データレコード、あるいは、1つ以上の前記冗長レコードを格納した記憶媒体とバッファを含む記憶装置であって、前記記憶媒体に格納した前記冗長レコードが属する前記パリティグループ内のデータレコードの値を、前記データレコードを格納した前記記憶媒体を含む記憶装置から受け取り、前記バッファに格納する手段を有することを特徴とする記憶装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、ディスクアレイ向きの高機能ディスク装置、および、ディスクアレイ向きの高機能ディスク装置と制御装置により構成される記憶装置サブシステムに関する。

【0002】

【従来の技術】発明に最も近い公知例として、以下に示

すPattersonの論文が知られている。

【0003】エー、シー、エム、シグモッド コンファレンス プロシーディング、1988年、6月、ページ109-116 (D. Patterson, et al: A Case for Redundant Array of Inexpensive Disks(RAID), ACM SIGMOD conference proceeding, Chicago, IL, June 1-3, 1988, pp. 109-11

6) Pattersonの論文は、ディスクアレイ上のデータ配置に関する技術を開示したものである。

【0004】ディスクアレイは、ディスクシステムの高性能化、高信頼化を実現するための機構である。ディスクアレイでは、高性能化のために、物理的には複数のディスク装置を、処理装置に対しては1台のディスク装置に見せかける。一方、高信頼化のためには、データを格納したディスク装置に障害が発生した場合、データの回復を行うための冗長データを別のディスク装置に格納しておく。

【0005】通常、ディスク装置のリード/ライト単位となるデータをレコードと呼ぶが、Pattersonの論文では、いくつかのレコード配置方法が提案されている。ただし、ディスクアレイを用いた場合、処理装置から見たリードライト単位であるレコードと、ディスク装置に実際に記録されるレコードとではデータ長が異なる場合がある。以下、前者を論理レコード、後者を物理レコードと呼ぶ。以下、Pattersonの論文で提案されているいくつかのレコード配置方法の説明を行う。

【0006】第1の配置方法は、論理レコード、すなわち、処理装置側から見たレコードを、ディスク装置上では、 m 個 ($m \geq 1$) の物理レコードに分割して格納する配置方法である。以下、この配置方法を、分割配置方法と呼ぶ。分割配置を用いた場合、1つの論理レコードを m 台のディスク装置との間で転送できることから、見かけ上データ転送速度を m 倍に向上させたのと同様の効果を得ることができる。次に、分割配置における冗長データの作成方法を説明する。分割配置では、論理レコードを分割した m 個の物理レコードに対し、 n 個 ($n \geq 1$) の冗長データが作成され、それぞれを、1つの物理レコード (全体で n 個ある) としてディスク装置に格納する。以下、処理装置が直接リード/ライトするデータを格納した物理レコードをデータレコード、冗長データを格納した物理レコードをパリティレコードと呼ぶ。また、 m 個のデータレコードと n 個のパリティレコードから構成されるグループを、パリティグループと呼ぶ。通常、パリティグループ内のパリティレコードの数が n 個であれば、 n 台までのディスク装置に障害が発生してもそのパリティグループのデータは回復可能である。第2の配置方法は、処理装置から見たリード/ライト単位である論理レコードを、1つの物理レコード、すなわち、1つのデータレコードとして、ディスク装置上に格納する配置方法である。以下、これを非分割配置と呼ぶ。し

たがって、論理レコードは、データレコードと等価なる。(それぞれの物理レコードには、データレコードあるいはパリティレコードが割り当てられるため、物理レコードと論理レコードは必ずしも等価にならない。すなわち、1つの論理レコードは、1つの物理レコードではあるが、1つの物理レコードは、1つの論理レコードであるというわけではないし、パリティレコードである場合もある。)非分割配置の特長は、ディスクアレイを構成するそれぞれのディスク装置ごとにリード/ライト処理が実行可能な点である。(分割配置方法をとると、リード/ライトのために複数のディスク装置を専有する必要がある。)したがって、非分割配置をとると、ディスクアレイ内で実行できるリード/ライト処理の多重度を向上させることが可能となり、性能向上を実現できる。非分割配置でも、m個のデータレコードから、n個のパリティレコードを作成し、ディスク装置に格納される。ただし、分割配置の場合、パリティグループ内のデータレコードの集合が、処理装置から見た1つの論理レコードを形成するのに対し、非分割配置の場合、データレコードのそれぞれが、処理装置から見るとまったく独立した論理レコードとなる。

【0007】一方、ディスクアレイではなく、一般のディスク装置のライト処理の高速化に関しては、ディスクキャッシュを利用した以下のような技術が開示されている。特開昭55-157053では、ディスクキャッシュを有する制御装置において、ライトアフト処理を利用してライト要求を高速化に関する技術が開示されている。具体的には、制御装置は、処理装置から受け付けたライトデータをキャッシュ内に書き込んだ段階で、ライト処理を完了させる。処理装置から受け付け、キャッシュ内に格納したデータのディスク装置への書き込みは、後から、制御装置のライトアフト処理によって実行される。

【0008】特開昭59-135563では、高信頼性を保証しながらライト処理を高速化する制御装置に関する技術が開示されている。

【0009】特開昭59-135563では、制御装置内にキャッシュメモリ以外に不揮発性メモリを有し、処理装置から受け取ったライトデータをキャッシュメモリと不揮発性メモリに格納する。ディスク装置へのライトデータの書き込みは、制御装置が、ライトアフト処理によって実行する。これにより、ライトアフト処理の高信頼化を図る。一方、特開昭60-114947では、2重書きディスク装置を制御するディスクキャッシュを有する制御装置に関する技術が開示されている。特開昭60-114947では、制御装置は、処理装置から受け取ったライト要求に対し、一方のディスク装置とキャッシュメモリに、処理装置から受け取ったライトデータを書き込む。もう一方のディスク装置には、制御装置が、処理装置からのリード/ライト要求とは非同期に、キャ

ッシュメモリに格納したライトデータを後から書き込む。制御装置が、処理装置からのリード/ライト要求とは非同期に、キャッシュメモリに格納したライトデータをディスク装置に後から書き込む動作をライトアフト処理と呼ぶ。

【0010】特開平2-37418では、2重書きディスク装置をディスクキャッシュを利用した高性能化に関する技術が開示されている。

【0011】特開平2-37418でも、制御装置内にキャッシュメモリ以外に不揮発性メモリを有し、処理装置から受け取ったライトデータをキャッシュメモリと不揮発性メモリに格納する。2つのディスク装置へのライトデータの書き込みは、制御装置が、ライトアフト処理によって実行する。

【0012】特開平3-37746では、ディスクキャッシュを有し、ライトアフト処理を実行する制御装置において、ライトアフト処理を効率よく実行することを目指すとしたディスクキャッシュ内のライトアフトデータの管理データ構造についての技術が開示されている。

【0013】

【発明が解決しようとする課題】ディスクアレイを用いた場合、処理装置からライト処理を受け付けた時、論理レコードの内容の変化に伴い、パリティレコードの内容も変更する必要が生ずる。したがって、(a)ライト対象となった論理レコードの更新値の転送処理以外に、

(b)パリティレコードの書き込みを行う転送処理と、

(c)パリティレコードの更新値を作成するのに必要な情報を揃えるための転送処理が、制御装置とディスク装置の間のデータ転送路に発生することになる。これらの転送処理は、ディスクアレイの適用により初めて生ずる処理であり、ディスクアレイ化により生ずる転送オーバーヘッドとなる。ライト処理に対する制御装置とディスク装置の間のデータ転送路のデータ転送増大量は、分割配置、非分割配置それぞれの場合で異なる。以下、それぞれの場合について、具体的に説明する。

【0014】分割配置の場合、処理装置から受け取ったライト対象とする論理レコードが、パリティグループ内のすべてのデータレコードの内容に相当することから、受け取った論理レコードの更新値から、パリティレコードが作成できる。このため、(c)のパリティレコードの更新値を作成するのに必要な情報を揃えるための転送処理は必要としない。以上より、制御装置とディスク装置の間のデータ転送路のデータ転送オーバーヘッドは、

(b)のパリティレコードの書き込みを行う転送処理の書き込み転送量だけである。

【0015】一方、非分散配置の場合、パリティレコードの更新値を作成するためには、(c)パリティレコードの更新値を作成するのに必要な情報を揃えるための転送処理として、以下の値の集合のうちのいずれかを得る処理が必要となる。

【0016】(1)ライト処理が発生した論理レコード(＝データレコード)の更新前の値、および、パリティレコードの更新前の値。

【0017】(2)ライト処理が発生した論理レコード(＝データレコード)が属するパリティグループ内の他のすべてのデータレコードの値。

【0018】通常、(1)に示した値を得る処理の方が、オーバーヘッドが小さいため、ライト処理が発生した場合、(1)に示した値を得る場合を前提にして、以下の説明を行う。(c)のパリティレコードの更新値を作成するのに必要な情報を揃えるための転送処理として、

(1)に示した値を得る処理を実行すると、パリティレコードを1つしか設けない場合($n=1$)でも、ライト処理が発生した論理レコード(＝データレコード)の更新前の値、および、パリティレコードの更新前の値と2回の転送が発生する。これ以外に、(a)ライト対象となった論理レコードの更新値の転送処理と、(b)パリティレコードの書き込みを行う転送処理が1回ずつ発生するため、制御装置とディスク装置のデータ転送回数は、計4回発生することになる。ディスクアレイを用いない場合、ライト処理では、発生するデータ転送処理は、(a)ライト対象となった論理レコードの更新値の転送処理の1回であるため、非分散配置のディスクアレイでは、制御装置とディスク装置の間のデータ転送量は、従来の4倍に増大する。

【0019】以上より、ディスクアレイの適用に伴い、制御装置が、処理装置との間で直接実行する転送処理に割り当てられる転送スループットは、上記の転送オーバーヘッドの分だけ低下することになる。

【0020】本発明の目的は、パリティレコードの操作に起因する、制御装置とディスク装置の間のデータ転送路の転送オーバーヘッドの増大を極力抑えることである。

【0021】

【課題を解決するための手段】以上述べた課題に対する本発明の目的を、いかに達成するかについて以下に述べる。

【0022】本発明では、パリティレコードの操作に起因する、制御装置とディスク装置の間のデータ転送路の転送オーバーヘッドを抑えることを目的とした基本機能として、ディスク装置にパリティレコード作成機能を設ける。ただし、単純に、ディスク装置にパリティレコード作成機能を設けた場合、非分散配置では、転送オーバーヘッドを削減できるが、分散配置では、転送オーバーヘッドを削減できない。以下、転送オーバーヘッドの最も少ないパリティレコードの数が1つの場合を例にして、その理由を説明する。

【0023】すでに述べたように、従来の動作では、非分散配置では、処理装置からのライト要求に伴い、データレコードの書き込み以外に、データレコードとパリティレコードの更新前の値の読みだし、パリティレコード

の書き込みという、従来の4倍の転送が発生する。一方、ディスク装置側にパリティ作成機能を設けた場合、制御装置側では、データレコードの更新前後の値からパリティレコードを作成するための中間値を作成し、これをディスク装置に転送する。中間値は、データレコードの更新前後の値の排他論理和から作成される。この中間値と更新前のパリティレコードの値から更新後のパリティレコードが作成される。ディスク装置側では、制御装置から受け取ったパリティレコードを作成するための中間値と、記録媒体から読みだしたパリティレコードの更新前の値から、パリティレコードの更新値を作成し、記録媒体に書き込む。以上のような動作では、制御装置とディスク装置の間の転送処理は、データレコードの読みだし、書き込み、および、パリティレコードを作成するための中間値の転送となり、ディスクアレイ適用以前の3倍までの転送量に抑えることができる。

【0024】しかし、分割配置の場合には、処理装置から受け取ったライト対象とする論理レコードの内容から、パリティレコードが作成できるため、制御装置が、パリティレコードを作成し、送信するのが最も効率的となる。したがって、ディスク装置側のパリティ作成機能を有効に活かすことができない。

【0025】さらに、本発明では、制御装置とディスク装置の間のデータ転送路の転送オーバーヘッドを抑えるために、基本機能であるディスク装置に設けたパリティレコード作成機能に、制御装置とディスク装置の間のブロードキャスト転送機能組み合わせる。この場合には、分散配置でも、非分散配置でも、転送オーバーヘッドを削減できる。以下、具体的な説明を行う。

【0026】まず、分散配置の場合について説明する。この場合、制御装置は、パリティグループに属する全ディスク装置に、論理レコードを分割せず、そのままの形で、ブロードキャストする。この時、論理レコードを受け取るパリティグループ内のディスク装置は、論理レコードの一部をデータレコードとして記録すべきディスク装置と、論理レコードに対応するパリティレコードを記録すべきディスク装置に分類できる。そのディスク装置が、データレコードを書き込むべきディスク装置であれば、論理レコードの中から自装置で書き込むべき部分をデータレコードとして取りだし、記録媒体に記録する。一方、そのディスク装置が、パリティレコードを書き込むべきディスク装置であれば、論理レコードからパリティレコードを作成して、記録媒体に記録する。

【0027】以上の機構では、制御装置が転送するのは論理レコードのみである。したがって、ディスクアレイの適用に伴い生ずる制御装置とディスク装置の間のデータ転送路の転送オーバーヘッドをなくすことが可能となる。

【0028】次に、非分散配置の場合について説明する。この場合も、制御装置は、パリティグループに属す

る全ディスク装置に、論理レコードを分割せず、そのままの形で、ブロードキャストする。この場合、論理レコードを受け取ったパリティグループ内のディスク装置は、論理レコード（＝データレコード）を記録媒体に記録すべきディスク装置、何も処理をしなくてよいディスク装置、および、論理レコードに対応するパリティレコードを記録すべきディスク装置に分類できる。パリティレコードを記録すべきディスク装置は、まず、論理レコードの更新前の値を、論理レコードを記録しているディスク装置から、転送してもらう。さらに、パリティレコードの更新前の値を記録媒体から読み出す。以上のようにして得た、パリティレコードの更新前の値、論理レコードの更新前の値、および、最初に受け取った論理レコードの更新値より、パリティレコードの更新値を作成し、記録媒体に書き込む。

【0029】一方、論理レコード（＝データレコード）として書き込むべきディスク装置は、パリティレコードを記録すべきディスク装置に、まず、論理レコード（＝データレコード）の更新前の値を送信する。この後、制御装置から受け取った論理レコードの更新値を記録媒体に記録する。以上の動作では、制御装置とディスク装置の間のデータ転送路を用いて転送されるデータは、論理レコードの更新値と更新前の値である。したがって、この場合、制御装置とディスク装置の間のデータ転送路の転送量を、ディスクアレイ適用以前の2倍までに転送量を抑えることができる。

【0030】

【作用】以下、本発明の作用について述べる。ただし、以下の説明では、転送オーバーヘッドの最も少ないパリティレコードの数が1つの場合を前提とする。

【0031】まず、非分散配置のディスクアレイに、基本機能であるディスク装置に設けたパリティレコード作成機能を適用した場合の作用について説明する。非分散配置では、本発明を適用しない場合、すでに述べたように、パリティレコードの更新値を作成するのに必要な情報を揃えるための転送処理として、ライト処理が発生した論理レコード（＝データレコード）の更新前の値、および、パリティレコードの更新前の値を転送すると、計4回のデータ転送処理が、制御装置とディスク装置の間で発生する。

【0032】一方、ディスク装置側にパリティ作成機能を設けた場合、制御装置側では、論理レコード（＝データレコード）の更新前と後の値からパリティレコードを作成するための中間値を作成し、これをディスク装置に転送する。ディスク装置側では、制御装置から受け取ったパリティレコードを作成するための中間値と、記録媒体から読み出したパリティレコードの更新前の値から、パリティレコードの更新値を作成し、記録媒体に書き込む。したがって、制御装置とディスク装置の間で実行されるデータ転送処理は、データレコードの更新前の値、

データレコードの更新値、および、パリティレコードの中間値の転送処理と、計3回の処理となる。本発明適用以前の転送回数は4回であるため、制御装置とディスク装置の間のデータ転送路の転送オーバーヘッドを削減するという目的を達成することができる。

【0033】次に、分散配置のディスクアレイに、基本機能であるディスク装置に設けたパリティレコード作成機能に、制御装置とディスク装置の間のブロードキャスト転送機能を組合せて適用した場合の作用について説明する。分散配置の場合、以上の機能を適用しないと、すでに述べたように、制御装置とディスク装置の間には、パリティレコードの転送がオーバーヘッドとして発生することになる。以上の機能を適用した場合、制御装置は、パリティグループに属する全ディスク装置に、論理レコードを分割せず、そのままの形で、ブロードキャストする。論理レコードを受け取った各ディスク装置は、以下のような処理を行う。そのディスク装置が、データレコードを書き込むべきディスク装置であれば、論理レコードの中から自装置で書き込むべき部分をデータレコードとして取りだし、記録媒体に記録する。一方、そのディスク装置が、パリティレコードを書き込むべきディスク装置であれば、論理レコードからパリティレコードを作成して、記録媒体に記録する。したがって、制御装置が転送するのは論理レコードのみで、パリティレコードの転送は必要なくなるため、制御装置とディスク装置の間のデータ転送路の転送オーバーヘッドを削減するという目的を達成することができる。

【0034】最後に、基本機能であるディスク装置に設けたパリティレコード作成機能に、制御装置とディスク装置の間のブロードキャスト転送機能を組合せ、非分散配置を適用した場合の作用について説明する。

【0035】この場合も、制御装置は、パリティグループに属する全ディスク装置に、論理レコードを分割せず、そのままの形で、ブロードキャストする。論理レコードの更新値を受け取ると、パリティレコードを記録すべきディスク装置は、まず、論理レコード（＝データレコード）の更新前の値を、論理レコードを記録しているディスク装置から、転送してもらう。さらに、パリティレコードの更新前の値を記録媒体から読み出す。以上のようにして得た、パリティレコードの更新前の値、論理レコードの更新前の値、最初に受け取った論理レコードの更新値より、パリティレコードの更新値を作成し、記録媒体に書き込む。

【0036】一方、論理レコード（＝データレコード）を書き込むべきディスク装置は、パリティレコードを記録すべきディスク装置に、まず、論理レコードの更新前の値を送信する。この後、制御装置から受け取った論理レコードの更新値を記録媒体に記録する。以上の動作では、制御装置とディスク装置の間で実行されるデータ転送処理は、データレコードの更新前の値、および、デー

タレコードの更新値と、計2回の処理となる。分散配置の場合、本発明適用以前の転送回数は4回であるため、制御装置とディスク装置間のデータ転送路の転送オーバーヘッドを削減するという目的を達成することができる。

【0037】

【実施例】以下、本発明の実施例を説明する。

【0038】まず、各実施例に共通する内容について説明する。図13は、本発明の対象となる計算機システムの構成である。計算機システムは、処理装置1300、制御装置1305、および、1台以上のディスク装置1304により構成する。処理装置1300は、CPU1301、主記憶1302、および、チャネル1303により構成される場合があってもよい。制御装置1305は、処理装置1300からのリード/ライト要求にしたがって、処理装置1300と、ディスク装置1304の間で、転送処理を実行する。制御装置バッファ1310は、制御装置1305がリード/ライトするデータを、一時的に蓄えるバッファである。ただし、図28に示したように、制御装置内に2つ以上のディレクタ1307を含み、それぞれのディレクタ1307が処理装置1300からリード/ライト要求を受け付け、リード/ライト動作を実行するような構成でも本発明は有効である。

【0039】図14は、本発明の対象となる別の計算機システムの構成である。図13に示した構成との差異は、制御装置1305が、キャッシュメモリ1308、ディレクトリ1309、不揮発性メモリ1400、および、不揮発性メモリ管理情報1401を含む点である。キャッシュメモリ（以下、単にキャッシュと略す。）1308は、ディスク装置1304の中のアクセス頻度の高いデータをロードしておく。ディレクトリ1309には、キャッシュ1308の管理情報を格納する。不揮発性メモリ1400は、不揮発の媒体であり、キャッシュ1308と同様に、ディスク装置1304の中のアクセス頻度の高いデータをロードしておく。不揮発性メモリ管理情報1401も不揮発の媒体であり、不揮発性メモリ1400の管理情報を格納する。この場合、制御装置1305は、処理装置1300からのリード/ライト要求とは、非同期に、ディスク装置205とキャッシュ1308との間で、リード/ライト動作を実行する。通常、処理装置1300がディスク装置との間で、リード/ライトするデータの単位は、レコードと呼ばれる。ディスクアレイを適用しない場合、処理装置1300から見たレコードと、ディスク装置1304上に格納されたレコードは、等しい。一方、ディスクアレイを適用した場合、処理装置1300から見たレコードと、ディスク装置1304上に格納されたレコードが、ディスクアレイのレコード配置によって異なる場合がある。以下、ディスクアレイを適用した場合のデータの記録形式について説明する。

【0040】次に、図15、図16を用いて、ディスクアレイの適用した場合のディスク装置1304上の記録形式について説明する。

【0041】図15に示すように、ディスク装置1304と制御装置1305の間で、リード/ライトされる単位、すなわち、ディスク装置1304に記録されている単位を、物理レコード1502と呼ぶ。本発明においては、ディスク装置1304上に格納されている物理レコード1502には、データレコード1500とパリティレコード1501の2種類が存在する。データレコード1500は、処理装置1300が、直接リード/ライトするデータを格納した物理レコード1502である。一方、パリティレコード1501は、ディスク装置1304に障害が発生し、データレコード1501の内容が消失した時、その消失した内容を回復する処理に用いるレコードである。この場合、データレコード1500の値が変更されると、これに対応して、パリティレコード1501の内容も変更する必要が生ずる。

【0042】図16は、ディスクアレイにおけるパリティグループ1600の構成である。ディスク装置a1601からディスク装置d1604までのm個のディスク装置1304上に、それぞれ対応するデータレコード1500が格納されている。これらのm個のデータレコード1500から、n個のパリティレコード1501が作成され、それぞれ対応するディスク装置e1605からディスク装置f1606に格納される。したがって、図16では、m個のデータレコード1500とn個のパリティレコード1501から、パリティグループ1600が構成されている。一般に、n個のパリティレコード1501を含むパリティグループ1600においては、そのパリティグループ1600内のレコード1502が格納されているm+n個ディスク装置のうち、n台のディスク装置1304が故障しても、パリティグループ1600内のすべてのレコード1502の内容を回復することができる。以上より、ディスクアレイを適用すると、ディスク装置1304の高信頼化が実現できる。

【0043】図16のパリティグループ1600においては、ディスク装置a1601からディスク装置c1603にデータレコード1500が、ディスク装置d1604からディスク装置e1605に格納されている。ただし、ディスク装置a1601からディスク装置c1603までのディスク装置1304上に格納するすべてのレコード1502を、データレコード1500とする必要はない。同様に、ディスク装置d1604からディスク装置e1605までのディスク装置1304上のすべてのレコード1502が、パリティレコード1500というわけではない。

【0044】また、図16においては、パリティグループ1600が、ディスク装置a1601からディスク装置e上に作成されているが、パリティグループ1600

が作成されるディスク装置1304の集合は、各パリティグループ1600が異なったディスク装置1304上に存在してもよい。例えば、別のパリティグループ1600がディスク装置b1602からディスク装置c1603に作成されてもよい。同様に、パリティグループ1600を構成するレコード1502の数も $m+n$ 個には限定されない。

【0045】ただし、以下の実施例においては、簡略化のため、パリティグループ1600の構成を、図16に示したように、 m 個のデータレコード1500と n 個のパリティレコード1501からなる構成とする。

【0046】ディスクアレイを適用した場合、処理装置1300から見たレコードと、ディスク装置1304上のデータレコード1500の関係が、レコードの配置方法により、異なる場合がある。以下、図17と図18を用いて、2つの代表的なディスクアレイのレコード配置方法である分散配置と非分散配置の説明を行う。

【0047】図17は、分散配置のディスクアレイのレコード配置を表す。以下の実施例では、処理装置1300と制御装置の間でリード/ライトされる単位、すなわち、処理装置から見たリード/ライト単位を論理レコード1700と呼ぶ。分散配置のディスクアレイでは、論理レコード1700は、 m 個に分割され、それぞれがデータレコード1500として、ディスク装置1304に格納される。さらに、 m 個の分割されたデータレコード1500から n 個のパリティレコード1501が作成され、それぞれディスク装置1304に格納される。したがって、分散配置の場合、1つの論理レコード1700が、1つのパリティグループ1600を構成する。

【0048】分割配置を用いた場合、1つの論理レコード1700を m 台のディスク装置との間で転送できることから、見かけ上データ転送速度を m 倍に向上させたのと同様の効果を得ることができる。ただし、1つのリード/ライト動作に対応して、少なくとも m 台のディスク装置1304を占有する必要が生ずる。

【0049】図18は、非分散配置のディスクアレイのレコード配置を表す。非分散配置では、1つの論理レコードがデータレコード1500としてディスク装置1304に格納される。すなわち、論理レコード1700とデータレコード1500は1対1の対応関係をもつ、すなわち、等価である。したがって、非分散配置では、1つのパリティグループ1600には、 m 個の論理レコード1700が含まれる。

【0050】非分割配置をとった場合、論理レコード1700のリード/ライトに1つのディスク装置1304だけを占有すればよいから、ディスクアレイ内で実行できるリード/ライト処理の多重度を向上させることが可能となる。

【0051】ディスクアレイの適用により、分散配置あるいは非分散配置のいずれの場合も、論理レコード17

00のライト処理に伴い、パリティレコード1501の内容も書き換える必要が生ずる。したがって、ディスク装置1304と制御装置1305の間で、リード/ライトされるデータ量が、ディスクアレイの適用に伴い、従来に比較し、増大するという問題が発生する。特に、非分散配置の場合、リード/ライトされるデータ量の増加率が高い。以下、それぞれの配置をとった場合のリード/ライトデータの増加量について説明する。

【0052】分割配置の場合、処理装置1300から受け取ったライト対象とする論理レコード1700を分割した m 個のデータレコード1500の内容から、パリティレコード1501が作成できる。また、論理レコード1700、すなわち、 m 個のデータレコード1500に相当するライトデータ量は、従来のディスク装置でもライトされるため、ライトデータ量の増加はない。したがって、分割配置の場合、ディスクアレイにおいてリード/ライトされるデータの増分は、パリティレコードのライトデータ量だけである。

【0053】一方、非分散配置では、処理装置1300から受け取ったライト対象とする論理レコード1700の内容だけからは、パリティレコード1501が作成できない。この場合、ライト処理が発生した論理レコード1700の更新値以外に、例えば、以下の値の集合のうちのいずれかを取得するためのリード処理を実行する必要がある。

【0054】(1)ライト処理が発生した論理レコード1700、すなわち、対応するデータレコード1500の更新前の値、および、パリティレコード1501の更新前の値

(2)ライト処理が発生した論理レコード1700(=データレコード1500)が属するパリティグループ1600内の他のすべてのデータレコード1500の値
通常、(1)に示した値を得る方が、オーバーヘッドが小さいため、ライト処理が発生した場合、(1)に示した値を得る方法をとった場合について説明する。したがって、パリティレコード1500を書き込むというライト処理に加えて、さらに、(1)に示した値を得るためのリード処理が、従来ディスク装置に比較し、転送オーバーヘッドとなる。このため、パリティレコード1500が1つしか設けない場合($n=1$)でも、ディスクアレイ適用以前に比較し、リード/ライトされるデータ量は、4倍に増大する。

【0055】本発明の概要は、ディスクアレイを適用した時、論理レコード1700のライト処理に伴い発生する、ディスク装置1304と制御装置1305の間のリード/ライトデータの転送量を、ディスク装置1304にパリティレコード501の作成機能を設けることにより削減するというものである。以下、図1を用いて、その内容を説明する。

【0056】まず、ディスク装置1304の構成を説明

する。ディスク装置1304内の記憶媒体101は、実際に物理レコード1501を記録した媒体である。プロセッサ100は、制御装置1304から記憶媒体101上の物理レコード1502をリード/ライトする。本発明では、さらに、プロセッサ100は、パリティレコード作成機能をもつ。バッファ102は、パリティレコード1501の更新値等を作成する際に利用するメモリである。バッファ管理情報107は、バッファ102上に、どのようなデータを格納したを示す管理情報である。

【0057】ディスク装置1304は、制御装置1305の指示にしたがい、パリティレコード1501の更新処理を実行する。以下、その内容を説明する。プロセッサ100内のパリティ情報入力部103は、制御装置1305、あるいは、記憶媒体101から、パリティレコード1501の更新値を作成するための情報を、パリティ入力データ108として、バッファ102に入力する(109)。パリティ作成部104は、パリティ情報入力部103によりバッファ102に格納されたパリティ入力データ108を入力し(110)、パリティレコード1501の更新値をパリティ更新値106として作成し、バッファ102に格納する(111)。この後、パリティレコードライト部105により、バッファ102に格納したパリティ更新値106を記憶媒体101に書き込む(112)。

【0058】以上が、本発明の概要であるが、以下の実施例では、ディスク装置1304に設けるパリティレコード501の作成機能に関し、3種類の実施例を開示する。まず、各実施例の概要を説明する。

【0059】第1の実施例は、単純に、ディスク装置1304にパリティレコード作成機能を設け、非分散配置をとったディスクアレイに関する。

【0060】以下、第1の実施例の概要を図2を用いて説明する。すでに述べたように、非分散配置をとるディスクアレイでは、論理レコード1700(=データレコード1500)の更新値、ライト処理が発生した論理レコード1700の更新前の値、および、パリティレコード1501の更新前の値により、パリティレコード1501の更新値を作成することができる。

【0061】ディスク装置1304にパリティレコード作成機能がなければ、非分散配置では、処理装置からのライト要求に伴い、制御装置1305とディスク装置1304の間で、論理レコード1700(=データレコード1500)とパリティレコード1501の読みだし、および、論理レコード1700とパリティレコード1501の書き込みという、ディスクアレイ適用以前の4倍の転送が発生する。

【0062】ディスク装置1304側にパリティ作成機能を設けた場合の動作については、図2を用いて説明する。制御装置側1305では、ディスク装置1304か

ら読み出した(206)論理レコード1700(=データレコード1500)の更新前の値である更新前データ209と、処理装置1300から受け取った(204)論理レコード1700の更新後の値である論理レコード更新値210をパリティ中間値作成部201に入力し

(212)、パリティレコードを作成するための中間値、すなわち、パリティ中間値200を作成する。さらに、制御装置1305は、作成したパリティ中間値200を、キャッシュ1308に格納する(213)。ただし、制御装置1305にキャッシュ1308が存在しない場合、制御装置バッファ1310に格納する。この後、制御装置1305は、パリティ中間値送信部202により、パリティ中間値200をディスク装置1304に転送する(214)。

【0063】ディスク装置1304側では、パリティ中間値入力部203により、制御装置1305から受け取ったパリティ中間値200をバッファ102に格納する(214)。次に、パリティ読み出し部205により、パリティレコード1501の更新前の値を、記憶媒体101から読みだし、更新前パリティ207として、バッファ102に格納する(215)。さらに、パリティ中間値200と更新前パリティ207をパリティ作成部a208に入力し(216)、パリティ更新値106を作成し、バッファ102に格納する(217)。この後、パリティレコードライト部105により、バッファ102に格納したパリティレコード1501の更新値を記憶媒体101に書き込む(112)。

【0064】以上のような動作では、制御装置1305とディスク装置1304の間の転送処理は、データレコード1500の読みだし、書き込み、および、パリティ中間値200の転送となり、ディスクアレイ適用以前の3倍までに、転送量に抑えることができる。

【0065】第2、第3の実施例では、ディスク装置1304に設けたパリティレコード作成機能と制御装置1305とディスク装置1304の間のブロードキャスト転送機能組み合わせて利用する。

【0066】第2の実施例は、ディスク装置1304に設けたパリティレコード作成機能と制御装置1305とディスク装置1304の間のブロードキャスト転送機能とを、組み合わせて、分散配置のディスクアレイに適用した場合の実施例である。以下、図3を用いてその概要を説明する。

【0067】この場合、制御装置1305は、論理レコードライト部300により、パリティグループ1600に属するm+n台のディスク装置1305に、処理装置1300から受け取った論理レコード更新値210を分割せず、そのままの形で、ブロードキャストする(304)。この時、論理レコード更新値210を受け取るパリティグループ1600内のディスク装置1304は、論理レコード更新値210の一部をデータレコード15

00の更新値として記録すべきm台のデータディスク装置301と、更新対象となる論理レコード1700に対応するパリティレコード1501を記録すべきn台のパリティディスク装置302に分類できる。ただし、データディスク装置301、および、パリティディスク装置302は、各ディスク装置1304に固定的な属性ではなく、転送対象となる論理レコード1700ごとに決定される属性である。

【0068】データディスク装置301の場合、まず、制御装置1305より受け取った論理レコード更新値210を、論理レコード入力部306により、バッファ102に格納する(304)。次に、データレコードライト部a303により、論理レコード更新値210の中から自装置で書き込むべきデータレコード1500の更新値を取りだし、記憶媒体101に書き込む(307)。

【0069】パリティディスク装置302の場合、論理レコード更新値210のバッファ102への格納(304)は、同様であるが、次にパリティ作成部b305により、自装置で書き込むべきパリティ更新値106を、論理レコード1700からを作成する(308)。この後、パリティレコードライト部105により、バッファ102に格納したパリティレコード1501の更新値を、記憶媒体101に書き込む(112)。

【0070】パリティレコード作成機能とブロードキャスト転送機能を用いず、分割配置のディスクアレイを適用すると、制御装置1305とディスク装置1304の間で転送されるデータの増分は、パリティレコード1501の転送量となる。しかし、パリティレコード作成機能とブロードキャスト転送機能を用いた場合、制御装置1305が転送するのは論理レコード1700のみであるため、ディスクアレイの適用に伴い生ずる制御装置1305とディスク装置1304の間のデータ転送路の転送オーバーヘッドをなくすることが可能となる。

【0071】図3に示した処理は、非分割配置のディスクアレイにおいても、パリティグループ1600内のすべての論理レコード1700(=データレコード1500)をまとめて更新する場合にも適用できる。というのは、図3に示した処理は、非分割配置のディスクアレイにおいて、制御装置1305が、パリティグループ1600内のすべてのデータレコード1500の更新値をブロードキャストするという処理に他ならないためである。

【0072】第3の実施例は、ディスク装置1304に設けたパリティレコード作成機能と制御装置1305とディスク装置1304の間のブロードキャスト転送機能とを組み合わせ、非分散配置のディスクアレイに適用した場合の実施例である。以下、その概要を図4を用いて説明する。

【0073】第3の実施例も、第2の実施例と同様に、論理レコードライト部300により、パリティグループ

1600に属するm+n台のディスク装置1304に、処理装置1300から受け取った論理レコード更新値210を、ブロードキャストする(304)。この場合、パリティグループ1600内のディスク装置1304は、論理レコード1700(=データレコード1500)をそのまま記録すべき1台のデータディスク装置301、何も処理をしなくてよいm-1台のディスク装置、および、論理レコード1700に対応するパリティレコード1501を記録すべきn台のパリティディスク装置302に分類できる。第2の実施例と同様、データディスク装置301、パリティディスク装置302、および、何もしなくてもよいディスク装置1304は、ディスク装置1304に固定的な属性ではなく、論理レコード1700対応に決定される属性である。また、何も処理をしなくてよいm-1台のディスク装置には、論理レコード更新値210を送らないようにしてもよい。なお、何も処理をしなくてよいm-1台のディスク装置は、特に説明すべき点がないため図4には示さなかった。

【0074】パリティディスク装置302は、まず、第2の実施例と同様、論理レコード入力部306により、論理レコード更新値210をバッファ102に格納する(304)。次に、データレコード受信部402により、論理レコード1700(=データレコード1500)の更新前の値である更新前データ209を、論理レコード1700(=データレコード1500)を記録しているデータディスク装置301から、受け取り、バッファ102に格納する(404)。さらに、パリティ読み出し部205により、更新前データ209を記憶媒体101から読み出す(215)。以上のようにして得た、論理レコード更新値210、更新前データ209及び更新前パリティ207をパリティ作成部c400に入力し、パリティ更新値106を作成する(405)。この後、パリティレコードライト部105により、バッファ102に格納したパリティレコード1501の更新値106を記憶媒体101に書き込む(112)。

【0075】データディスク装置301の場合も、まず、第2の実施例と同様、論理レコード入力部306により、論理レコード更新値210をバッファ102に格納する。次に、n台のパリティディスク装置302に、更新前データ209を送信する。この後、データレコードライト部b401により、制御装置1305から受け取った論理レコード更新値210をデータレコード1500の更新値として、記憶媒体101に書き込む。

【0076】以上の動作では、制御装置1305とディスク装置1304の間のデータ転送路を用いて転送されるデータは、論理レコード1700の更新前後の値である。したがって、この場合、制御装置1305とディスク装置1304の間のデータ転送量を、ディスクアレイの適用以前の2倍までに抑えることができる。

【0077】以下、各実施例の詳細を説明する。まず、第1の実施例の詳細について説明する。

【0078】第1の実施例は、ディスク装置1304にパリティレコード作成機能を設け、非分散配置をとったディスクアレイに関する。第1の実施例において、制御装置1305とディスク装置1304の間のデータ転送路の転送量の削減を可能にするのは、制御装置1305が、処理装置1300から受け取った論理レコード1700（＝データレコード1500）の更新値と、論理レコード1700の更新前の値、すなわち、更新前データ209から、パリティ中間値200を作成し、ディスク装置1304に転送する点である。

【0079】以下、以上の機能を実行する図2に示したパリティ中間値作成部201とパリティ中間値送信部202の処理フローについて説明する。以上の処理は、処理装置1300から受け取った論理レコード1700を更新するためのライト要求と同期して実行してもよいし、あるいは、ライト要求の完了報告を返した後、非同期に実行してもよい。ただし、同期、非同期にかかわらず、少なくとも処理装置1300から受け取った論理レコード1700を、制御装置メモリ1310、あるいは、キャッシュ1308に格納した後、実行を、開始する必要がある。

【0080】図2に示したパリティ中間値作成部201の処理フローを図5に示す。図5は、開始点aと開始点bの2つの開始点をもつ。制御装置1305は、まず、開始点aから実行を開始する。ステップ500で、ディレクトリ1309を参照して、キャッシュ1308に、更新前データ209が格納されているかをチェックする。この時、制御装置1305が、キャッシュ1308をもたず、制御装置メモリ1310をもつ場合、直ちに、ステップ501の実行にはいる。存在する場合、ステップ503へジャンプする。存在しない場合、ステップ501で、ライト対象となる論理レコード1700に対応するデータレコード1500を格納したディスク装置1304に、位置付け要求を発行し、一度処理を終了する。

【0081】開始点bは、ディスク装置1304の位置付け処理が完了した時、実行を開始するポイントである。ステップ502では、制御装置1304は、ディスク装置1304から、更新前データ209を、キャッシュ1308、あるいは、制御装置メモリ1310にロードする。キャッシュ1308にロードした場合、ディレクトリ1309の内容を更新する。

【0082】ステップ503では、制御装置1305は、キャッシュ1308、あるいは、制御装置メモリ1310に格納した論理レコード1700の更新値と更新前データ209より、パリティ中間値200を作成し、キャッシュ1308、あるいは、制御装置メモリ1310に格納する。

【0083】図30は、パリティ中間値送信部202の処理フローである。ステップ3000では、制御装置1305は、n台存在するパリティレコードを更新すべきディスク装置の内、まだパリティ中間値200を送っていない1台に、パリティ中間値200を送信する。ステップ3001では、n台すべてに送ったかどうかをチェックする。そうであれば、処理を終了し、そうでなければステップ3000に制御を戻す。

【0084】次に、図2に示したディスク装置1305側の各処理部の処理フローを説明する。

【0085】まず、図6を用いて、パリティ中間値入力部203の処理フローを説明する。ステップ600で、プロセッサ100は、制御装置1305から受け取ったパリティ中間値200をバッファ102に格納する。この時、バッファ管理情報107を、パリティ中間値200のバッファ102への格納に対応して、更新する。次に、図7を用いて、パリティ読み出し部205の処理フローを説明する。まず、プロセッサ100は、ステップ700で、バッファ管理情報107を参照し、更新前パリティ207が、バッファ102に存在するかを、チェックする。存在する場合、処理を終了し、存在しない場合、ステップ701の実行に入る。バッファ102の容量が小さく、更新前パリティ207の存在が期待できない場合、直ちにステップ701から実行を開始してもよい。ステップ701では、記憶媒体101の位置付け処理が完了するのをまつ。位置付け処理が完了した時、プロセッサ100は、ステップ702で、更新前パリティ207をバッファ102に格納する。この時、バッファ管理情報107を、更新前パリティ207のバッファ102への格納に対応して、更新する。この後、処理を終了する。

【0086】図8は、パリティ作成部a208の処理フローである。ステップ800で、プロセッサ100は、パリティ中間値200と更新前パリティ207からパリティ更新値106を作成し、バッファ102に格納する。この時、図1に示すバッファ管理情報107を、パリティ更新値106のバッファ102への格納に対応して、更新する。この後、処理を終了する。

【0087】図9は、パリティレコードライト部105の処理フローである。まず、ステップ800で、プロセッサ100は、記憶媒体101の位置付け処理が完了するのをまつ。位置付け処理が完了した時、プロセッサ100は、ステップ801で、バッファ102に格納したパリティレコード1501の更新値を記憶媒体101に書き込む。この後、処理を終了する。

【0088】次に、第2の実施例の詳細について説明する。

【0089】第2の実施例は、ディスク装置1304に設けたパリティレコード作成機能と制御装置1305とディスク装置1304の間のブロードキャスト転送機能

とを組み合わせ、分散配置のディスクアレイに適用した場合の実施例である。以下、図3に示した制御装置1304とデータディスク装置301とパリティディスク装置302内の各処理部の内容を処理フローを用いて説明する。

【0090】図10は、制御装置1305内の図3に示す論理レコードライト部300の処理フローである。本処理は、処理装置1300から受け取った論理レコード1700を更新するためのライト要求と同期して実行してもよいし、あるいは、ライト要求の完了報告を返した後、非同期に実行してもよい。この場合、制御装置1305は、ステップ1000で、パリティグループ1600に属する $m+n$ 台のディスク装置1305に、処理装置1300から受け取った論理レコード1700を分割せず、そのままの形で、ブロードキャストする。

【0091】次に、データディスク装置301の各処理部について説明する。

【0092】図11は、論理レコード入力部306の処理フローである。プロセッサ100は、ステップ1100で、制御装置205より受け取った論理レコード1700を、バッファ102に格納する。この時、バッファ管理情報107を、論理レコード1700のバッファ102への格納に対応して、更新する。この後、処理を終了する。

【0093】図12は、データレコードライト部a303の処理フローである。まず、ステップ1200で、プロセッサ100は、記憶媒体101の位置付け処理が完了するのをまつ。位置付け処理が完了した時、プロセッサ100は、ステップ1201で、論理レコード1700の中から自装置で書き込むべき部分をデータレコード1500として取りだし、記憶媒体101に書き込む。この後、処理を終了する。

【0094】最後に、パリティディスク装置302の各処理部について説明する。

【0095】論理レコード入力部306の処理フローは、データディスク装置301の場合と共通であるため、説明を省略する。

【0096】図19は、図3に示すパリティ作成部b305の処理フローである。ステップ1900で、プロセッサ100は、自装置で書き込むべきパリティ更新値106を、論理レコード1700からを作成し、バッファ102に格納する。この時、バッファ管理情報107を、パリティ更新値106のバッファ102への格納に対応して、更新する。この後、処理を終了する。

【0097】パリティレコードライト部105の処理フローは、第1の実施例と共通であるため、説明を省略する。

【0098】最後に、第3の実施例の詳細について説明する。

【0099】第2の実施例は、ディスク装置1304に

設けたパリティレコード作成機能と制御装置1305とディスク装置1304の間のブロードキャスト転送機能と、分散配置組み合わせで利用した場合の実施例である。以下、図4に示した制御装置1304とデータディスク装置301とパリティディスク装置302内の各処理部の内容を処理フローを用いて説明する。ただし、制御装置1305内の論理レコードライト部300の処理フローは、第2の実施例と共通であるため、説明を省略する。

【0100】まず、パリティディスク装置302の各処理部について説明する。ただし、論理レコード入力部306の処理フローは、第2の実施例と共通であるため、説明を省略する。

【0101】図20と図21は、データレコード受信部a402の処理フローである。プロセッサ100は、図20の処理フローから実行を開始する。ステップ2000で、バッファ管理情報107を参照して、データレコード1500の更新前の値である更新前データ209が、バッファ107内に存在するかをチェックする。存在する場合は、ステップ2001で、更新前データ209を格納したデータディスク装置301に、更新前データ209の送信要求を発行する。この後、一度処理を終了する。バッファ102の容量が少なく、更新前データ209の存在が、期待できない場合、直ちに、ステップ2001を実行してもよい。存在する場合、プロセッサ100は、ステップ2002で、更新前データ209を格納したデータディスク装置301に、更新前データ209の送信が必要ないことを通知する。この後、処理を終了する。

【0102】図21の処理フローは、更新前データ209を格納したデータディスク装置301から、更新前データ209を送られてきた時、実行を開始する処理を表した処理フローである。ステップ2100で、プロセッサ100は、更新前データ209を、バッファ102に格納する。この時、バッファ管理情報107を、更新前データ209の格納に対応して、更新する。この後、処理を終了する。

【0103】パリティ入力部205の処理フローは、第1の実施例と共通であるため、説明を省略する。

【0104】図22は、パリティ作成部c400の処理フローである。プロセッサ100は、ステップ2200で、論理レコード更新値210、更新前データ209および更新前パリティ207より、パリティ更新値106を作成し、バッファ102に格納する。この時、バッファ管理情報107を、パリティ更新値106のバッファ102への格納に対応して、更新する。この後、処理を終了する。

【0105】パリティレコードライト部105の処理フローは、第1の実施例と共通であるため、説明を省略する。

【0106】次に、データディスク装置301の各処理部について説明する。ただし、論理レコード入力部306の処理フローは、第2の実施例と共通であるため、説明を省略する。

【0107】図23は、データレコード送信部a403の処理フローである。ステップ2300で、プロセッサ100は、n台のパリティディスク装置302から、更新前データ209の送信要求がくるまで待つ。n台のパリティディスク装置302からの要求が到着すると、ステップ2301で、プロセッサ100は、n台すべてのパリティディスク装置302が、更新前データ209の送信を必要としないかをチェックする。そうなら処理を終了する。少なくとも1台のパリティディスク装置302が、更新前データ209の送信を必要とするなら、ステップ2302で、プロセッサ100は、バッファ管理情報107を参照して、更新前データ209が、バッファ102内に存在するかをチェックする。存在する場合は、ステップ2305へジャンプする。バッファ102の容量が少なく、更新前データ209の存在が、期待できない場合、直ちに、ステップ2302をスキップして、直接ステップ2303を実行してもよい。存在しない場合、記憶媒体101の位置付け処理が完了するのをまつ。位置付け処理が完了した時、プロセッサ100は、ステップ2304で、更新前データ209をバッファ102に格納する。この時、バッファ管理情報107を、更新前データ209のバッファ102への格納に対応して、更新する。

【0108】ステップ2305では、プロセッサ100は、n台のパリティディスク装置302のうち更新前データ209を必要とするパリティディスク装置302に、更新前データ209を、ブロードキャストする。この後、処理を終了する。ここで、ステップ2305で、すべてのパリティディスク装置302に、更新前データ209の送信し、更新前データ209が存在するパリティディスク装置302は、送られてきた更新前データ209を捨てるようにしてもよい。

【0109】図24は、データレコードライト部b401の処理フローである。まず、ステップ2400で、プロセッサ100は、記憶媒体101の位置付け処理が完了するのをまつ。位置付け処理が完了した時、プロセッサ100は、ステップ2401で、により、制御装置1305から受け取った論理レコード1700をデータレコード1500として、記憶媒体101に、書き込む。この後、処理を終了する。

【0110】何もしなくてよいm-1台のディスク装置1304については、特に説明すべきことはないため、説明は行わない。

【0111】第3の実施例では、制御装置1305とディスク装置1304の間のデータ転送路を用いて転送されるデータは、最大限、論理レコード1700の更新値

と更新前データ209である。しかし、ステップ2000から明らかのように、更新前データ209が、パリティディスク装置302のバッファ101内に存在すれば、更新前データ209は、制御装置1305とディスク装置1304の間のデータ転送路を用いて転送する必要はない。通常、論理レコード1700が、処理装置1300からライト対象になる直前に、処理装置1300からリードされることが多い。したがって、データディスク装置301が、処理装置1300からリード要求を受けた論理レコード1700（＝データレコード1500）を制御装置1305に送るとき、同時に、パリティディスク装置302にもブロードキャストしておくことが有効となる。というのは、論理レコード1700が、ライト対象となった時、更新前データ209が、パリティディスク装置302のバッファ101内に存在する確率を高めることができるためである。あるいは、無条件にブロードキャストを行うのではなく、処理装置1300から、制御装置1305経由で、当該リード要求でリードされた論理レコード1700は、この後、ライトされる確率が高いという情報を受け、この時だけ、論理レコード1700を、パリティディスク装置302にブロードキャストするようにしてもよい。

【0112】図25は、データディスク装置301が、論理レコード1700（＝データレコード1500）を制御装置1305に送るとき、同時に、パリティディスク装置302にもブロードキャストする動作を表している。

【0113】データディスク装置301内のデータレコード送信部b2500は、制御装置1305から、データレコード1500の送信要求を受けたとき、論理レコード1700（＝データレコード1500）を、制御装置1305とパリティディスク装置302にもブロードキャストする（2503）。ただし、データレコード送信部b2500が、上記のブロードキャストを行うのは、論理レコード1700の送信要求を受けた時、常に行うのではなく、その論理レコード1700が後でライトされる確率が高いという指示を、制御装置1305から通知された時だけでもよい。

【0114】制御装置1305内の、論理レコードリード部2501は、データディスク装置301から送られてきた論理レコード1700（＝データレコード1500）を、処理装置1700に送る。

【0115】パリティディスク装置302内の、データレコード入力部a2502では、送られてきた論理レコード1700（＝データレコード1500）を、更新前データ209としてバッファ102に格納する。

【0116】以下、各処理部の詳細を処理フローを用いて詳細に説明する。

【0117】図26は、データディスク装置301内のデータレコード送信部b2500の処理フローである。

ステップ2600で、プロセッサ100は、バッファ管理情報107を参照して、更新前データ209が、バッファ102内に存在するかをチェックする。存在する場合は、ステップ2305へジャンプする。バッファ102の容量が少なく、更新前データ209の存在が、期待できない場合、直ちに、直接ステップ2601を実行してもよい。存在しない場合、ステップ2601で、記憶媒体101の位置付け処理が完了するのをまつ。位置付け処理が完了した時、プロセッサ100は、ステップ2602で、更新前データ209をバッファ102に格納する。この時、バッファ管理情報107を、更新前データ209のバッファ102への格納に対応して、更新する。

【0118】ステップ2603では、プロセッサ100は、制御装置1305とn台のパリティディスク装置302に、更新前データ209を、ブロードキャストする。この後、処理を終了する。

【0119】図27は、制御装置1305内の、論理レコードリード部2501の処理フローである。まず、ステップ2700で、プロセッサ100は、データディスク装置301から送られてきた論理レコード1700（＝データレコード1500）を、処理装置1700に送る。この時、送られてきた論理レコード1700をキャッシュ1308に格納してもよい。キャッシュ1308に格納した場合、これに対応し、ディレクトリ1309を更新する。

【0120】図29は、パリティディスク装置302内の、データレコード受信部b2502の処理フローである。ステップ2900で、プロセッサ100は、受け取った論理レコード1700（＝データレコード1500）を、更新前データ209として、バッファ102に格納する。この時、バッファ管理情報107を、更新前データ209の格納に対応して、更新する。この後、処理を終了する。

【0121】

【発明の効果】ディスクアレイを用いた場合、処理装置からライト処理を受け付けた場合、論理レコードの内容の変化に伴い、パリティレコードの内容も書き換える必要が生ずる。したがって、ライト対象となった論理レコードの更新値以外に、パリティレコードの更新値を作成するのに必要な情報を揃えるための転送処理と、パリティレコードの書き込みを行う転送処理が、制御装置とディスク装置の間のデータ転送路に発生することになる。本発明では、ディスク装置にパリティレコード作成機能を設けることにより、パリティレコードの操作に起因する、制御装置とディスク装置の間のデータ転送路の転送オーバーヘッドを抑えることを可能にする。

【図面の簡単な説明】

【図1】本発明の概要

【図2】第1の実施例の概要

【図3】第2の実施例の概要

【図4】第3の実施例の概要

【図5】パリティ中間値送信部202の処理フロー

【図6】パリティ中間値入力部203の処理フロー

【図7】パリティ読み出し部205の処理フロー

【図8】パリティ作成部a208の処理フロー

【図9】パリティレコードライト部105の処理フロー

【図10】論理レコードライト部300の処理フロー

【図11】論理レコード入力部306の処理フロー

【図12】図12は、データレコードライト部a303の処理フロー

【図13】本発明の対象となる計算機システムの構成

【図14】本発明の対象となる別の計算機システムの構成

【図15】ディスク装置1304に記録されている物理レコード1502の構成

【図16】ディスクアレイにおけるパリティグループ1600の構成

【図17】分散配置のディスクアレイのレコード配置

【図18】非分散配置のディスクアレイのレコード配置

【図19】パリティ作成部b305の処理フロー

【図20】データレコード入力部a400の処理フロー

【図21】データレコード入力部a400の処理フロー

【図22】パリティ作成部c402の処理フロー

【図23】データレコード送信部a402の処理フロー

【図24】データレコードライト部b403の処理フロー

【図25】データディスク装置301が、論理レコード1700（＝データレコード1500）を制御装置1305に送るとき、同時に、パリティディスク装置302にもブロードキャストする動作を表す

【図26】データレコード送信部b2500の処理フロー

【図27】論理レコードリード部2501の処理フロー

【図28】制御装置内に2つ以上のディレクタ1307を含む構成

【図29】データレコード入力部b2502の処理フロー

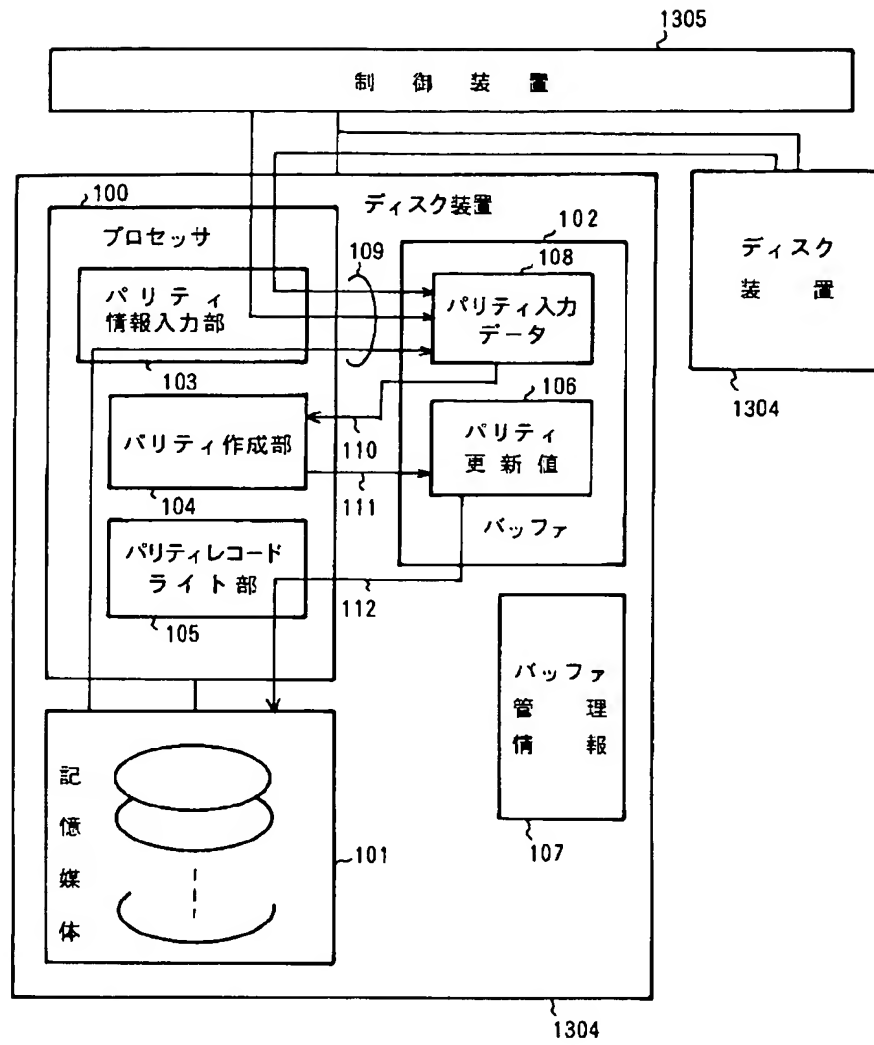
【図30】パリティ中間値送信部202の処理フロー

【符号の説明】

103…パリティ情報入力部、104…パリティ作成部、105…パリティライト部、102…バッファ、107…バッファ管理情報。

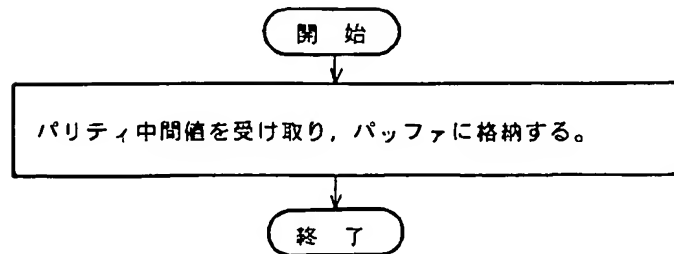
【図1】

図 1



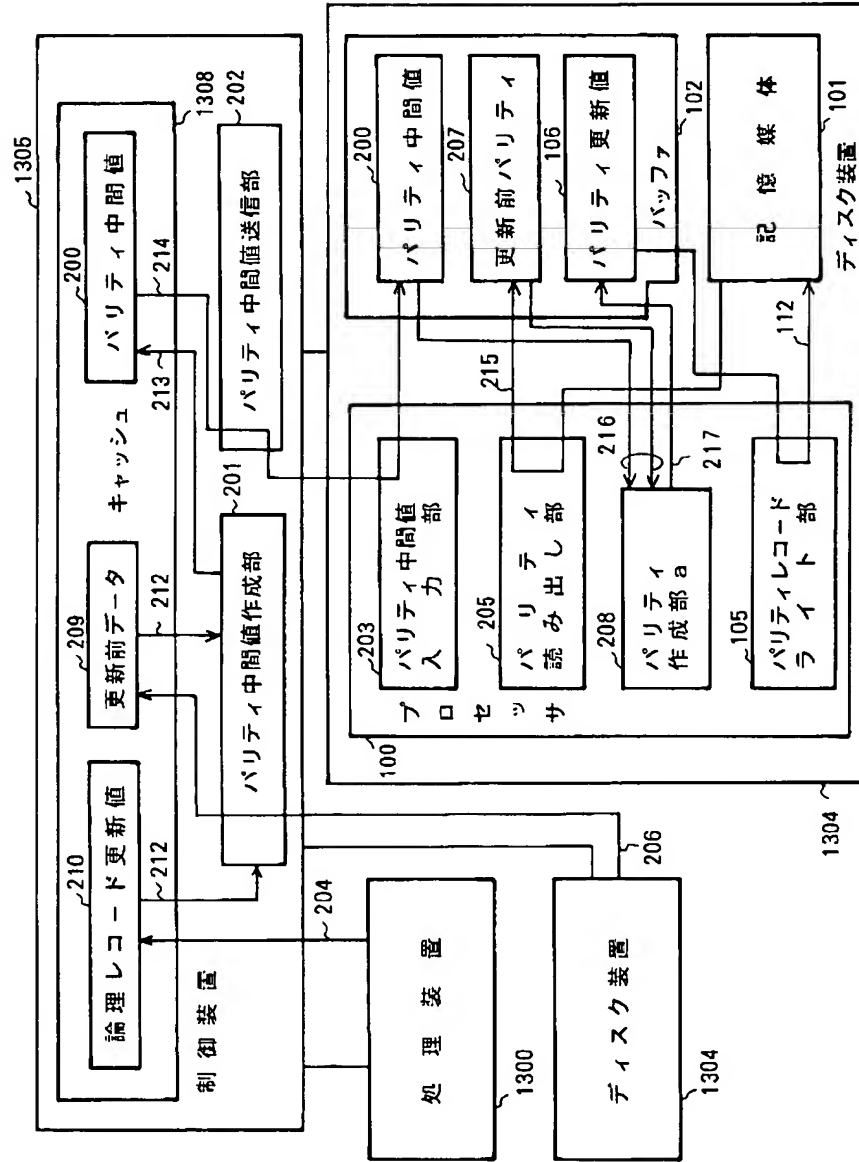
【図6】

図 6



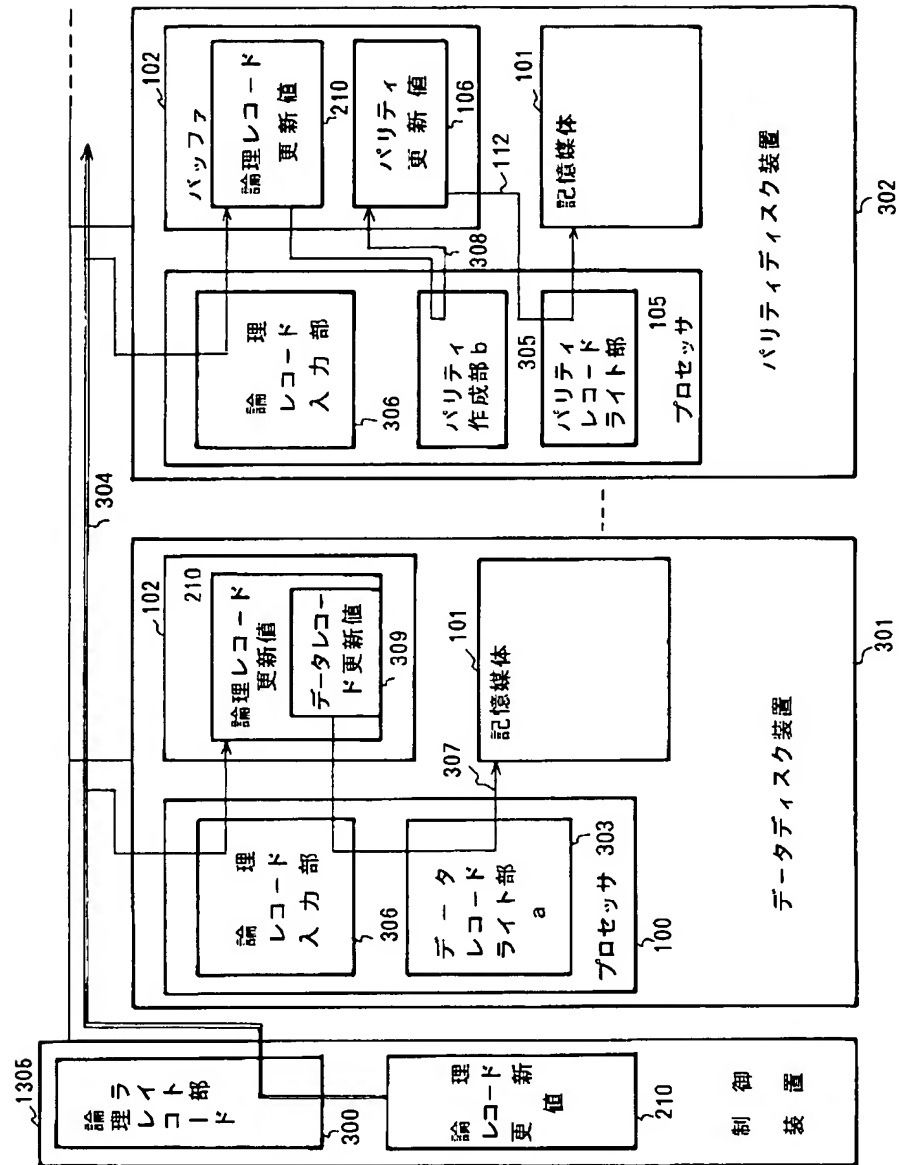
【図2】

図 2



【図3】

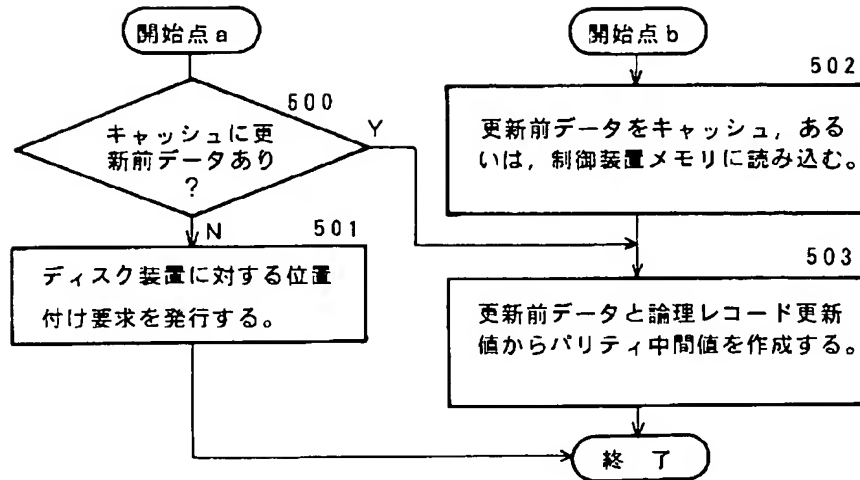
図 3





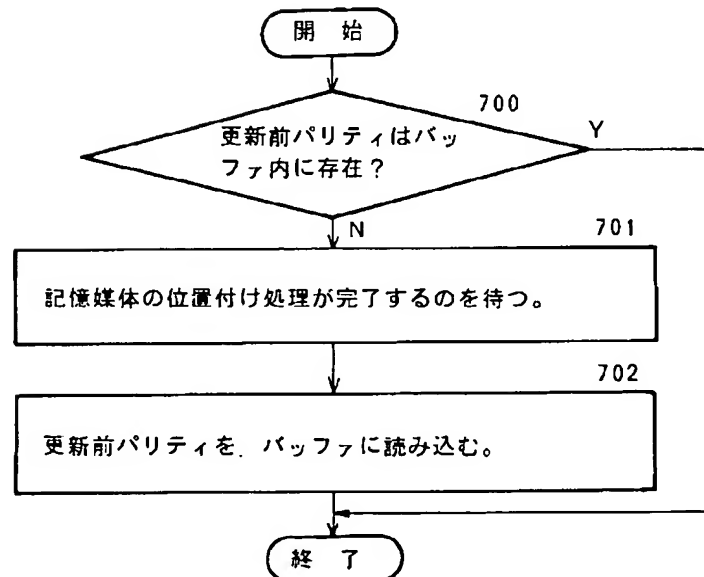
【図5】

図5



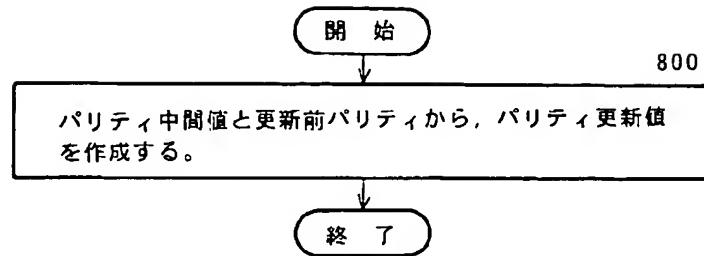
【図7】

図7



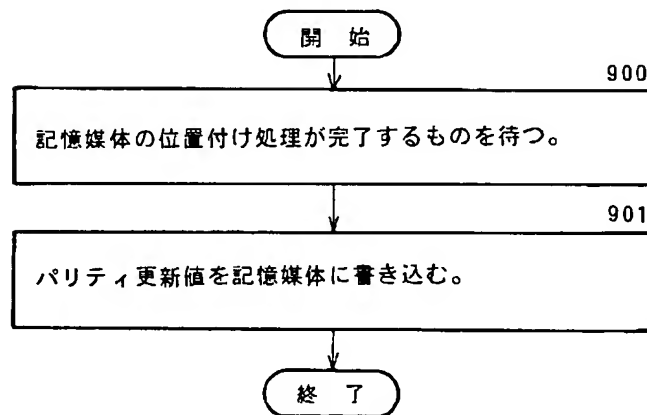
【図8】

図 8



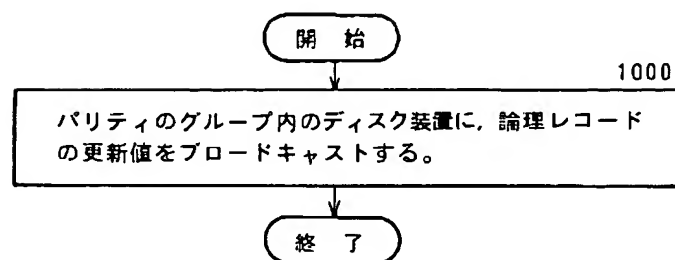
【図9】

図 9



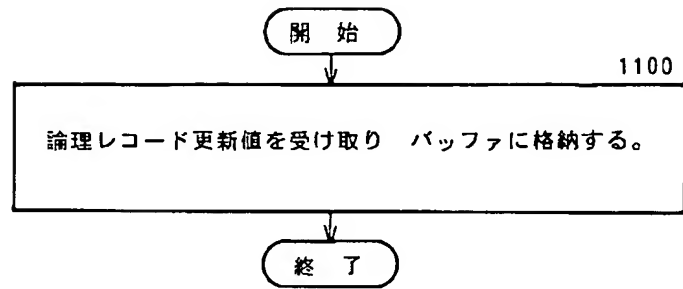
【図10】

図 10



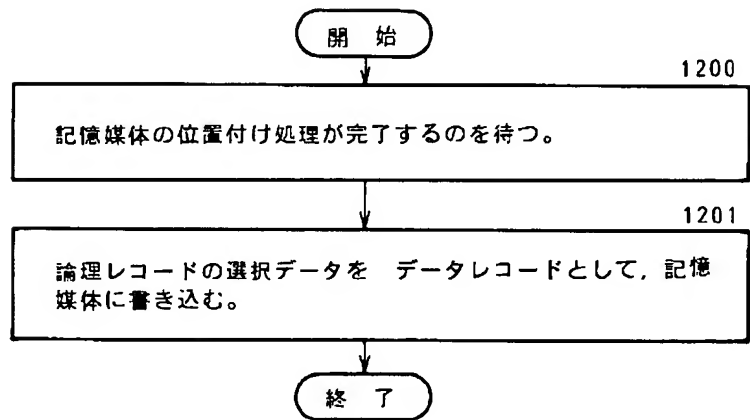
【図11】

図11



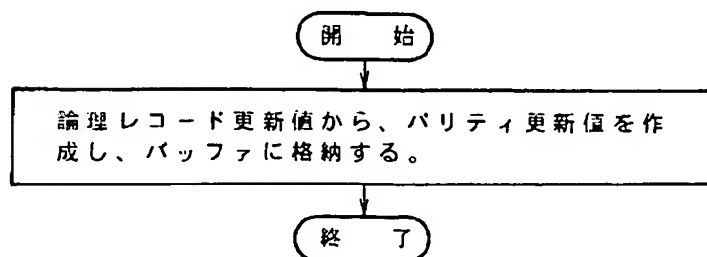
【図12】

図12



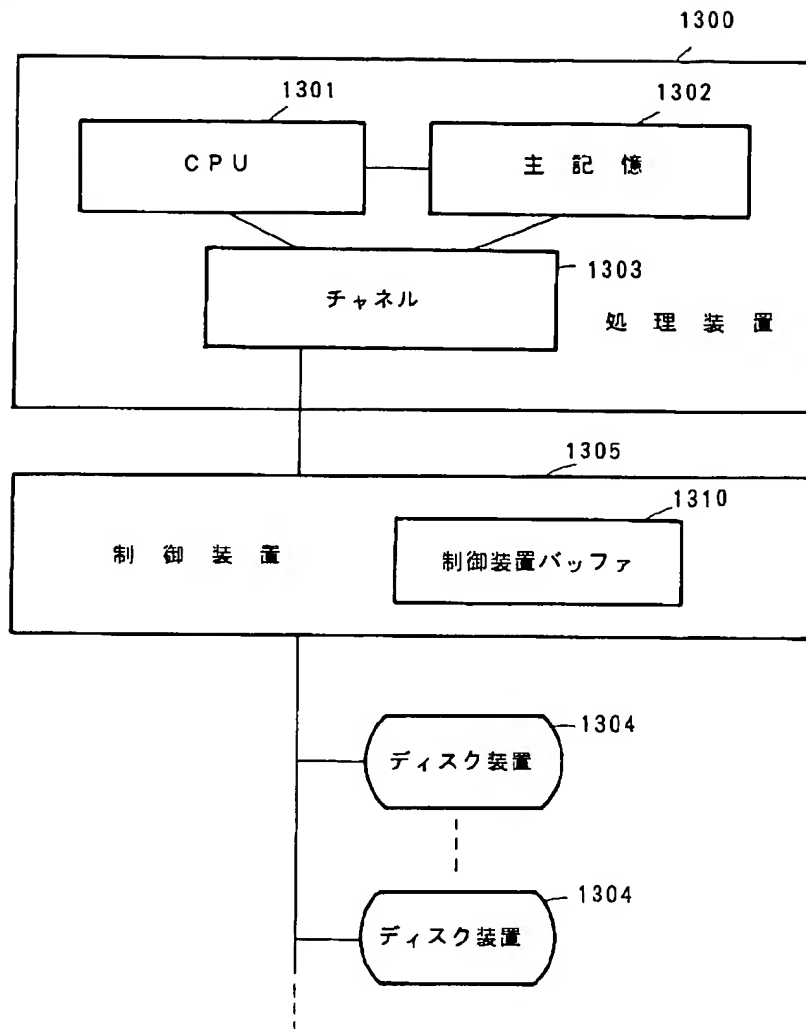
【図19】

図19



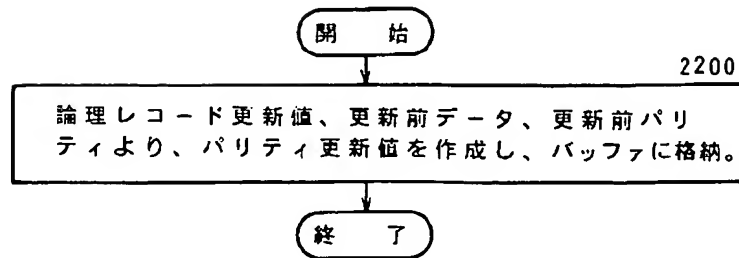
【図13】

図13



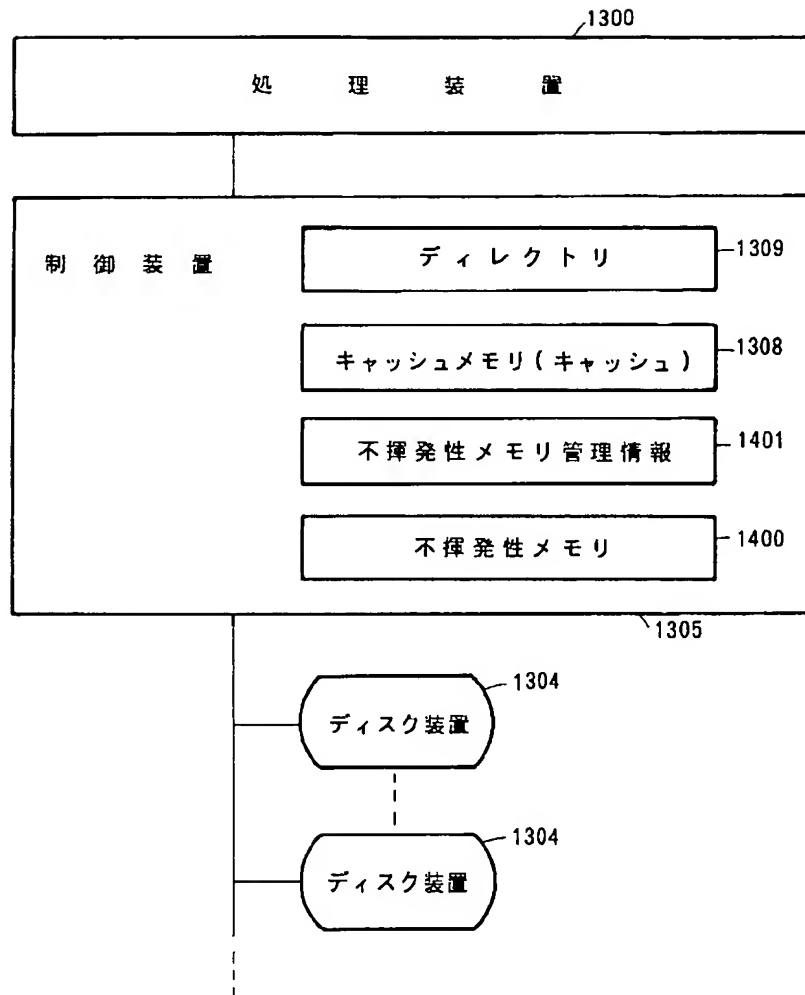
【図22】

図22



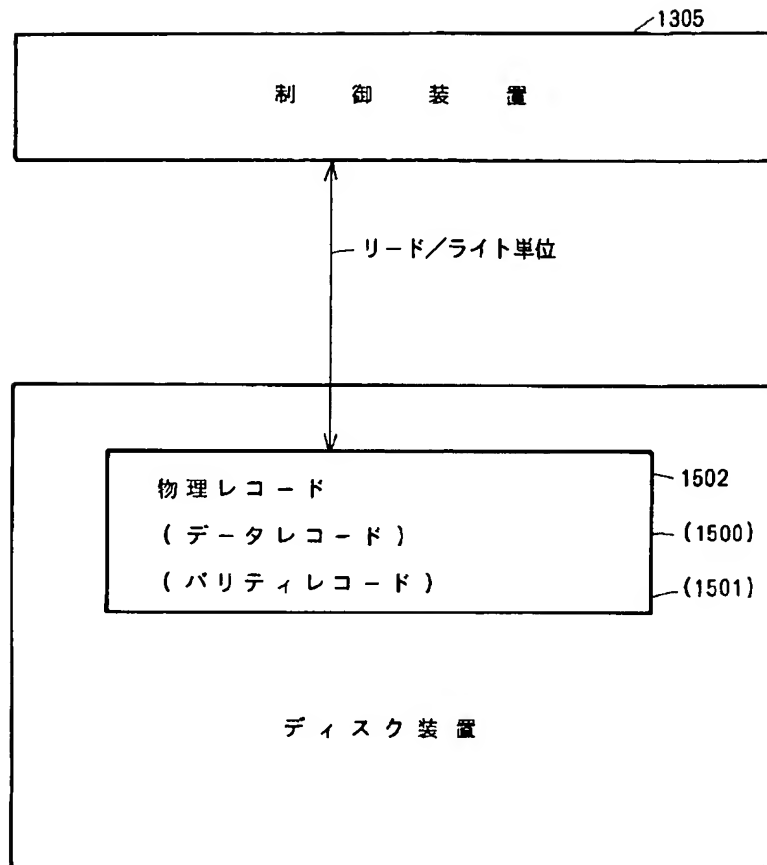
【図14】

図14



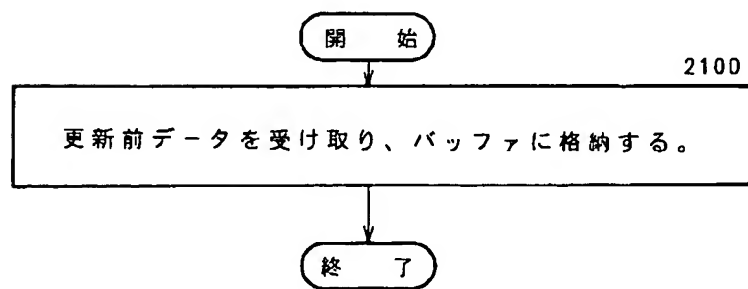
【図15】

図15



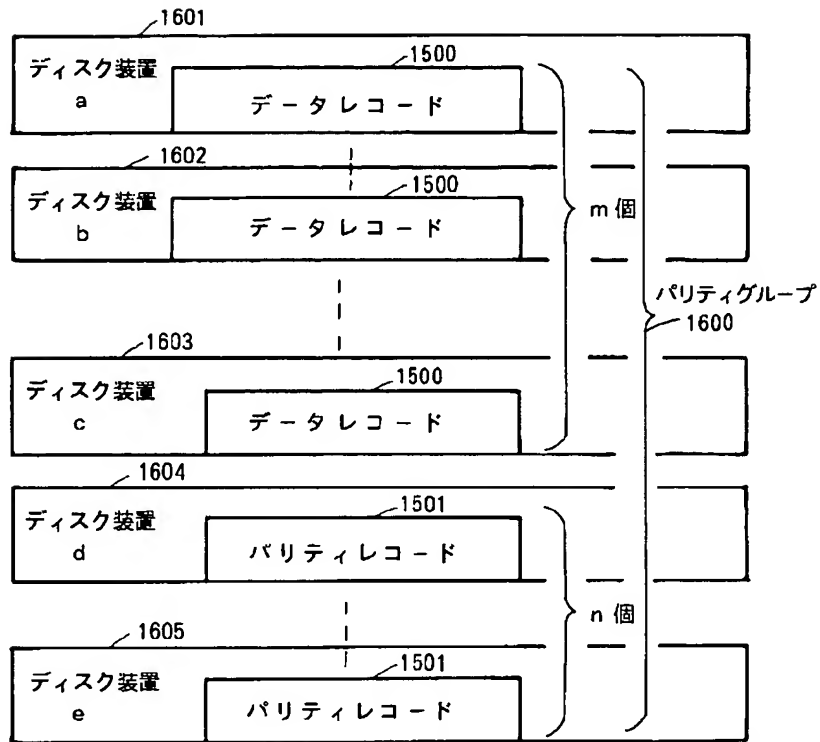
【図21】

図21



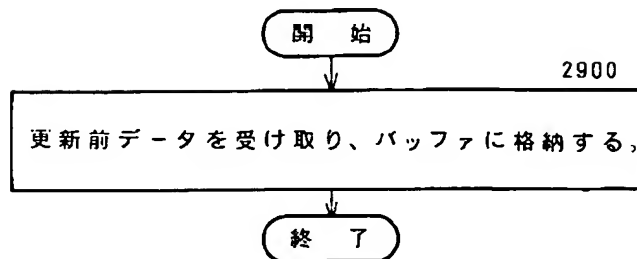
【図16】

図16



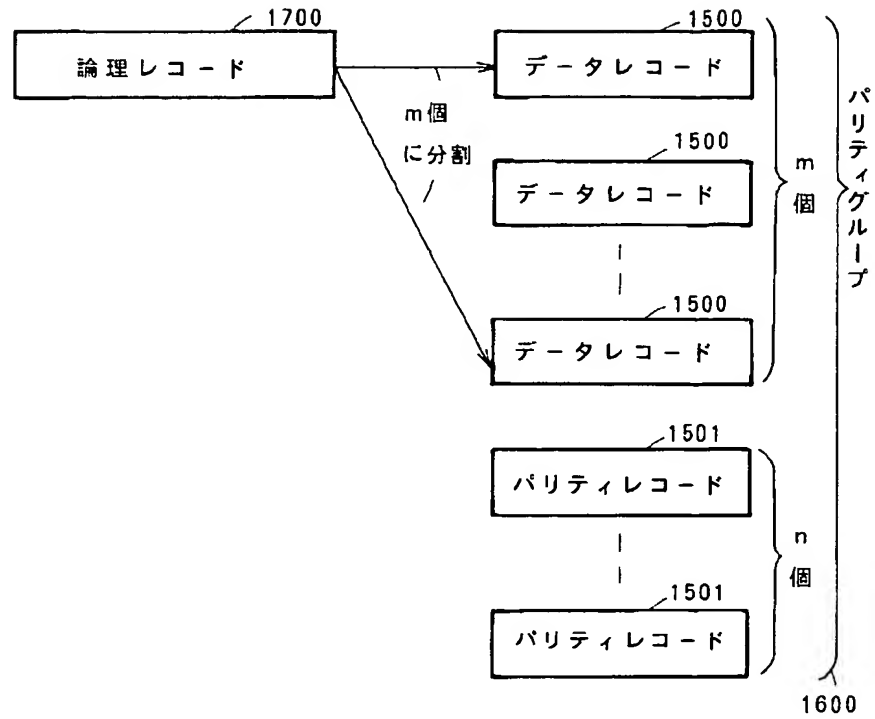
【図29】

図29



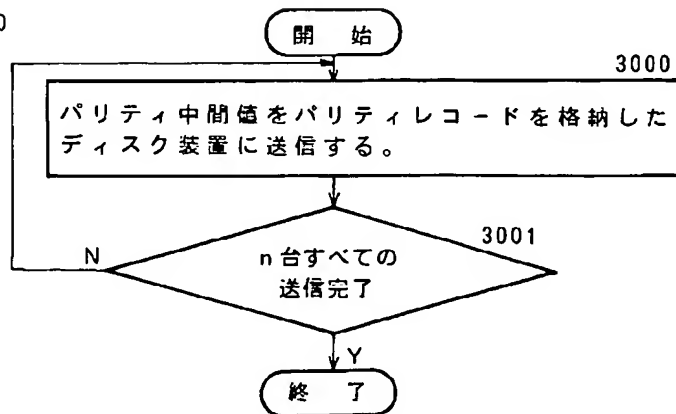
【図17】

図17



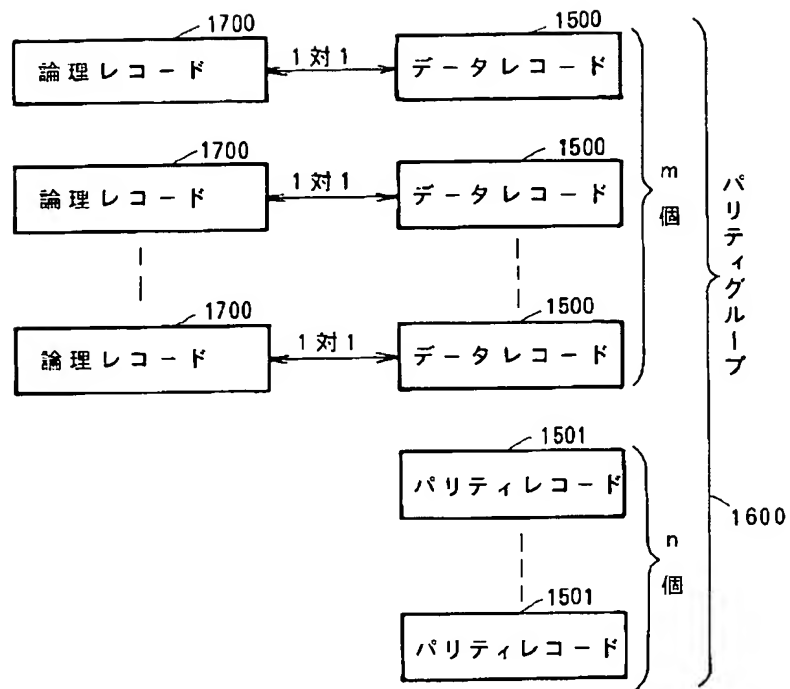
【図30】

図30



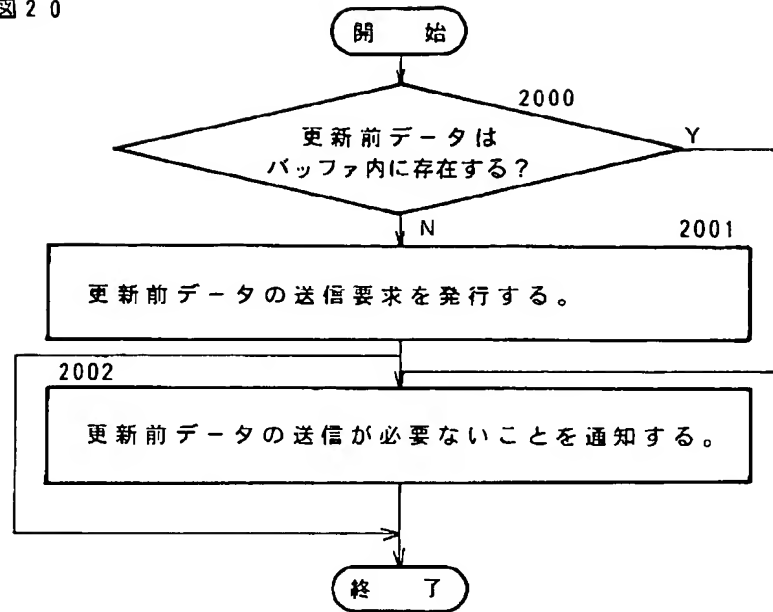
【図18】

図18



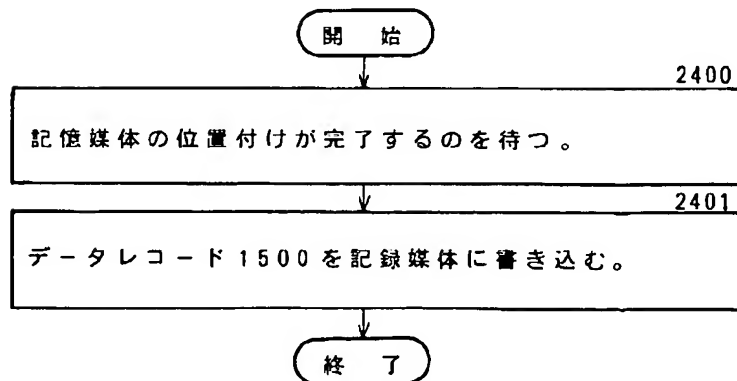
【図20】

図20



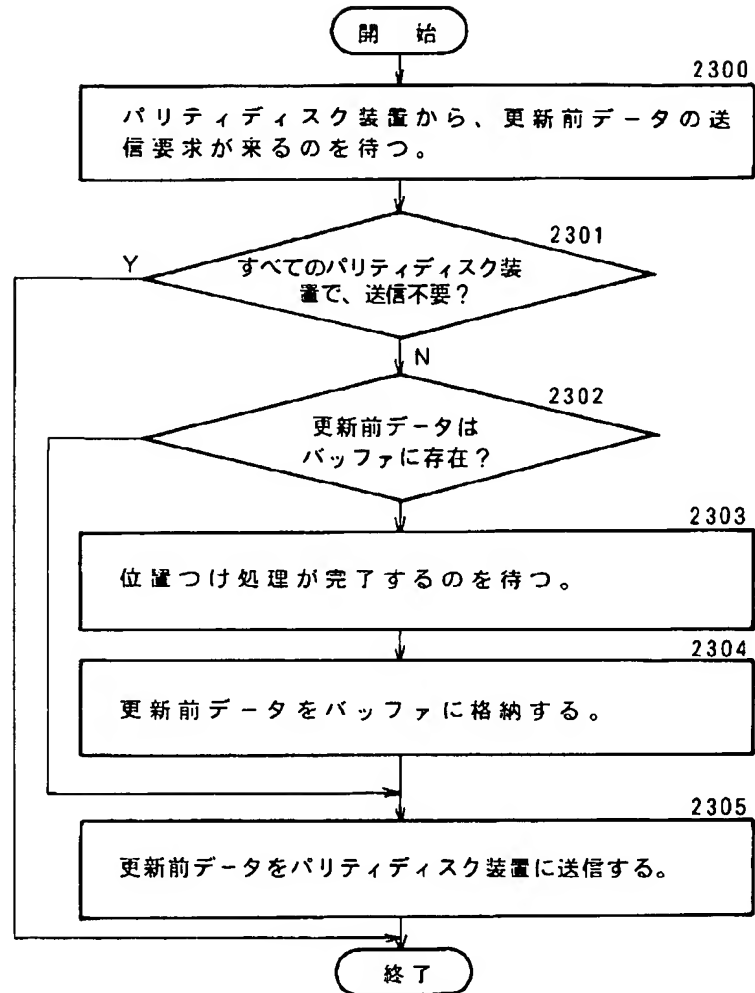
【図24】

図24



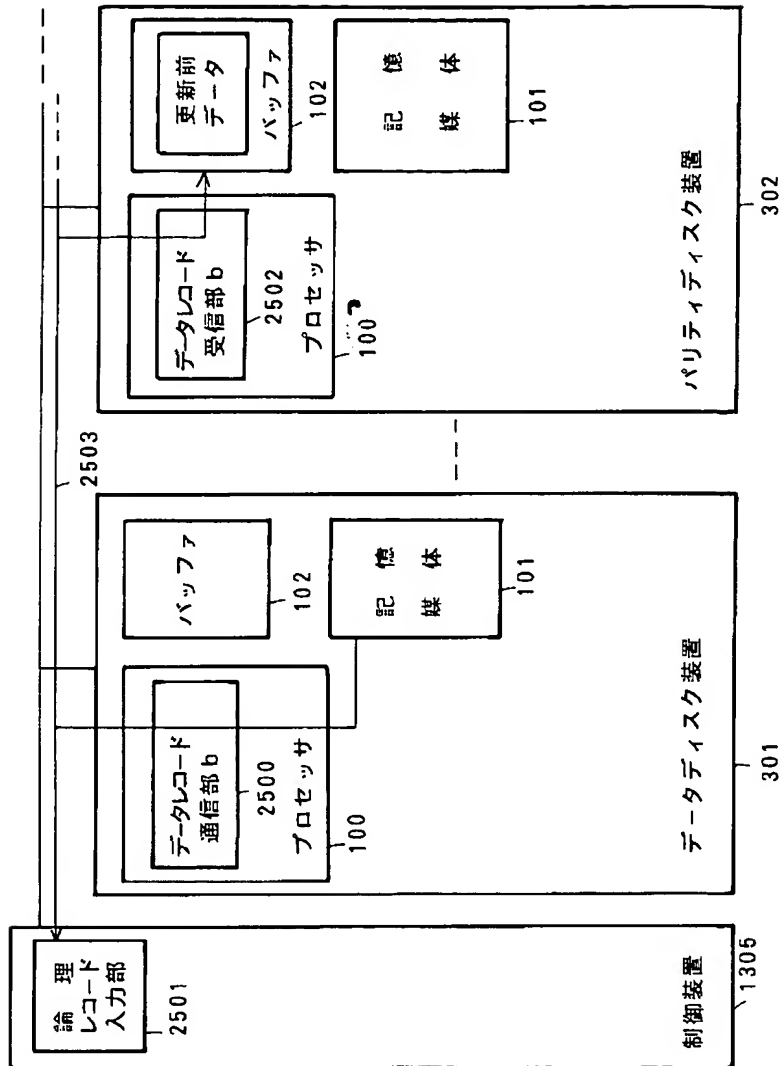
【図 23】

図 23



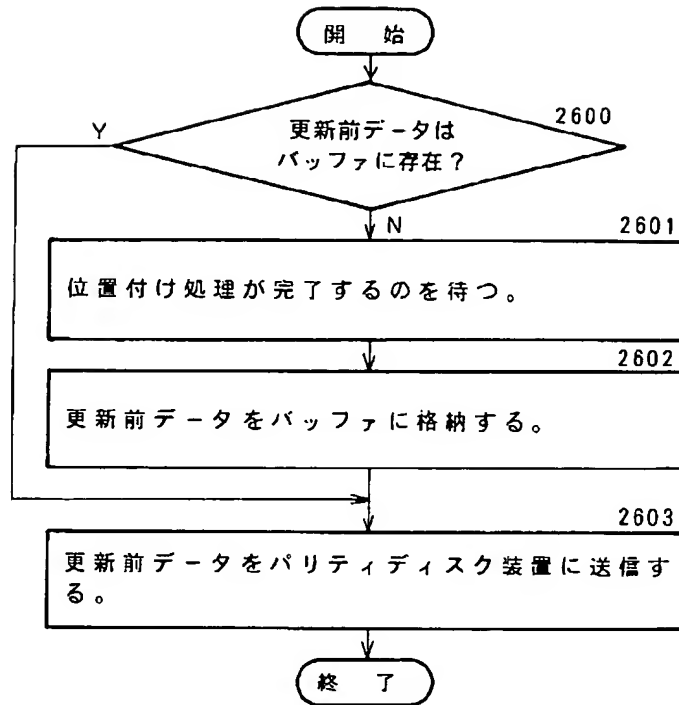
【図25】

図 25



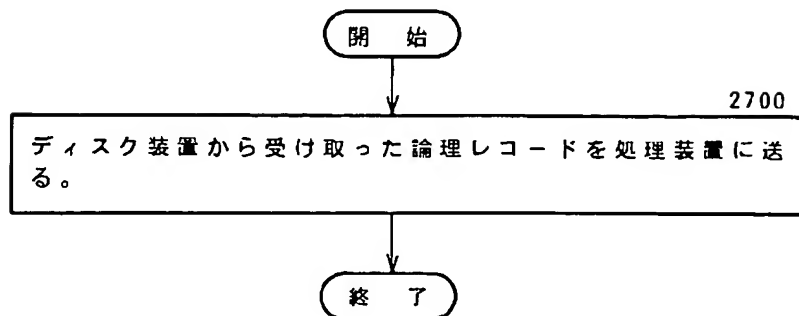
【図26】

図26



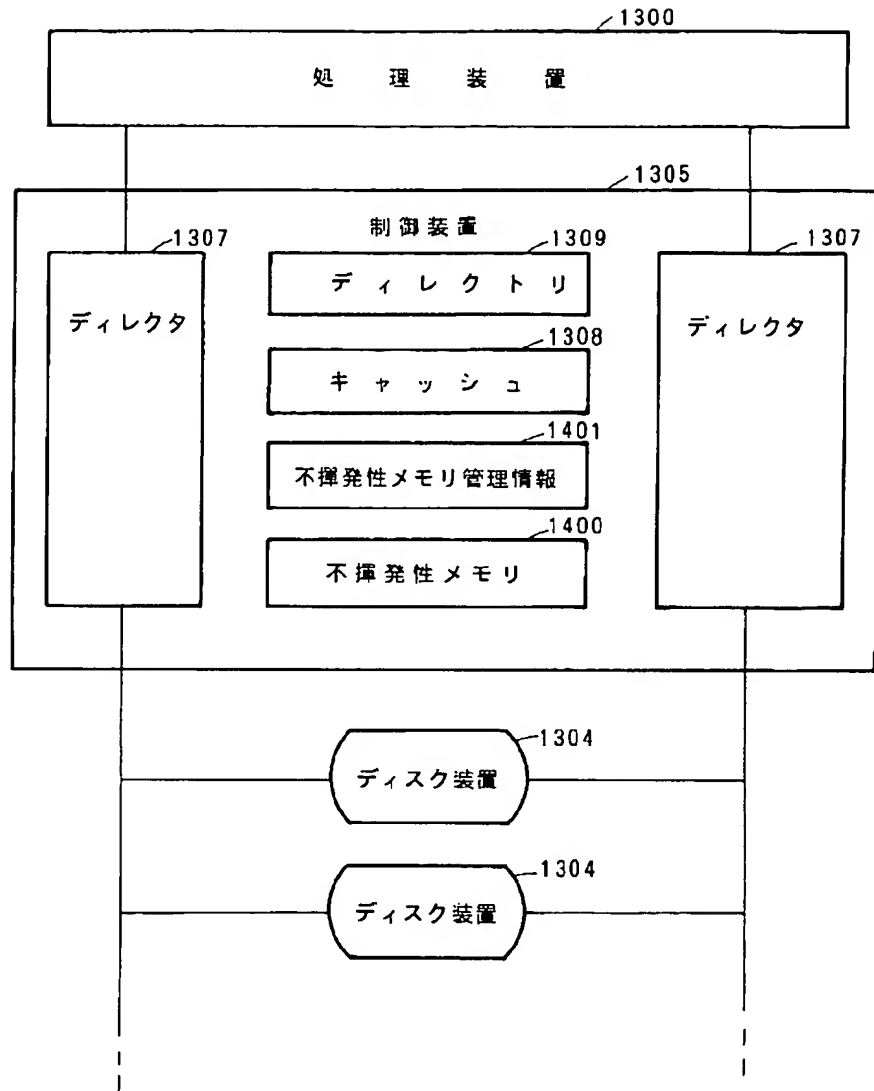
【図27】

図27



【図28】

図28



フロントページの続き

(72)発明者 倉野 昭
神奈川県小田原市国府津2880番地 株式会
社日立製作所小田原工場内

(72)発明者 猪股 宏文
神奈川県川崎市麻生区王禅寺1099番地 株
式会社日立製作所システム開発研究所内

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.